AUTOMATED FUNCTIONAL ASSESSMENT

OF SMART HOME RESIDENTS

By

PRAFULLA NATH DAWADI

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
School of Electrical Engineering and Computer Science

MAY 2015

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of PRAFULLA NATH DAWADI find it satisfactory and recommend that it be accepted.

Diane J. Cook, Ph.D., Chair

Lawrence B. Holder, Ph.D.

Maureen Schmitter-Edgecombe, Ph.D.

ACKNOWLEDGMENTS

Foremost, I would like to thank my advisor Dr. Diane J. Cook for introducing me to the amazing world of research in the machine learning and smart environments. I would like to thank her for providing me the warm atmosphere where I enjoyed learning every single day. Her guidance, valuable feedback, constant encouragement, and enthusiasm always fueled my research. It was truly a great learning experience while working with you. You are a wonderful teacher!

This dissertation would have never been completed without insights from Dr. Maureen Schmitter-Edgecombe. I would like to first thank her for answering all my questions about clinical neuropsychology and guiding me in each step of my research. I would like to acknowledge the diligent work of Dr. Maureen Schmitter-Edgecombe and her clinical neuropsychology team for meticulously designing and rigorously conducting the clinical studies that are an integral part of this dissertation. I would also like to thank Dr. Lawrence B. Holder for teaching me my first Machine Learning course and instilling my passion for Machine Learning. I would also like thank him for providing valuable feedback on the dissertation.

Finally, thanks to my colleagues who constantly supported me and cheered me up. Thanks to Gina, Joel, Haque, Yibo, Jess, Barnan, Ehsan, Shirin, Jason,

Anthony, Chris, Aaron and to all the CASAS lab members for their companionship and camaraderie.

AUTOMATED FUNCTIONAL ASSESSMENT OF SMART HOME RESIDENTS

Abstract

by Prafulla Nath Dawadi, Ph.D.
Washington State University
May 2015

Chair: Diane J. Cook

This dissertation proposes smart home-based intelligent techniques that perform automated assessment of a resident's well-being by monitoring their behavior inside the home. We hypothesize that the everyday behavior of smart home residents can be estimated by tracking residents' activities using smart home sensors and that machine learning algorithms can predict their cognitive and physical health utilizing behavioral information.

We first describe a cross-sectional study where we compare behavior differences across an entire population sample to assess activity quality and the individual's cognitive health. For this study, we introduce a machine learning-based framework for assessing the quality of eight different activities and one complex activity with interweaved sub-activities, called the Day Out Task. We compare our automated

assessment of task quality with direct observation scores and assess the ability of machine learning techniques to classify an individual's cognitive health using the same machine learning-based framework.

We then describe a longitudinal study where we use an individual as their own baseline to identify behavioral changes that can predict cognitive health and mobility. We first introduce a Clinical Assessment using Activity Behavior (CAAB) approach to model a smart home resident's daily behavior and predict the corresponding standard clinical assessment scores utilizing longitudinal smart home sensor data. CAAB extracts statistical features that describe characteristics of a resident's daily activity performance and trains the machine learning algorithms to predict the standard clinical assessment scores. We then introduce an activity curve to represent an abstraction of an individual's normal daily routine based on automatically recognized activities. We develop algorithms to detect changes in behavioral routines by comparing activity curves and use these changes to analyze the possibility of changes in cognitive or physical health. We evaluate all of our algorithms using real-world longitudinal smart home sensor data. We conclude that it is possible to assess the health and well-being of a smart home resident utilizing smart home sensor data and machine learning algorithms.

Contents

List of Tables

List of Figures

Dedication

To my father, Nabaraj Dawadi, and

my mother, Bindu Dawadi.

CHAPTER 1. INTRODUCTION

The Alzheimer's Association estimates that nearly 5.2 million Americans, one in eight people who are 65 or older, have Alzheimer's dementia and related dementia-like diseases [6]. Providing health care and support for these groups of older adults calls for huge health care and economic resources. Furthermore, the Alzheimer's Association estimates that treatment costs for individuals with such diseases will increase from $183 billion in 2011 to $1.1 trillion in 2050 (in 2011 dollars). Additionally, such diseases affect caregivers both emotionally and financially. The Association estimates that 70% of people with Alzheimer's live in their homes. The above-mentioned estimates, the economic impact, and the expected number of people that will be impacted are expected to massively increase in the next few decades. Therefore, we need to develop technologies that help older adults "age in place", meaning that the functionality will help senior residents live longer in their own homes than before [6, 118, 119, 120].

The long-term goal of this project is to design smart home technologies to offer functionalities that enable senior residents to age in place. The smart homes help older adults live independently for a longer time by providing following three main functionalities [28, 119]:

1. By assuring them that they are safe and they complete the activities of daily living.

2. By compensating for their sensory and cognitive impairments, for example, by reminding them to take medications.

3. By assessing their cognitive and physical health, for example, by detecting early indications of cognitive and physical decline.

Given the functionalities that smart homes offer, the possibilities of using these technologies for monitoring and assisting older adults are perceived as extraordinary and timely given the aging of population [101]. In this dissertation, we focus on developing the third smart home functionality listed above. We design algorithms to perform cognitive and physical assessment of a resident by monitoring everyday behavior using smart home technologies. Below we explain why smart home systems can effectively provide such functionality.

1. Smart homes technologies continuously monitor the everyday abilities of residents in their own home environment without governing, changing, or manipulating the individual's daily routines. Therefore, the smart home sensor data is ecologically valid [130].

2. Smart home sensor data provides clinicians and caregivers with rich behavioral information that is very difficult to gather otherwise. For example, sensor data

reflects behavioral information such as when a resident goes to bed and the number of times they wake up to go to the bathroom.

There are two main motivations for the dissertation. First, clinicians and caregivers consider the individual's ability to complete activities of daily living (or everyday behavior) such as cooking, eating, and sleeping as an important behavioral construct. Understanding everyday behavior helps clinicians identify difficulties a patient is having in real life, so that they can recommend measures to overcome the difficulties. Therefore, we develop machine learning algorithms that enhance the understanding of a person's everyday behavior by utilizing smart home sensor data.

The second motivation is that changes in the everyday behavior are frequently associated with decline in cognitive or physical health. Therefore, we develop algorithms to understand the cognitive and physical health of a smart home resident by continuously monitoring everyday behavior utilizing smart home sensor data.

The primary goal of this dissertation is to propose machine learning-based algorithms to perform automated cognitive and physical health assessment by monitoring everyday behavior using smart home sensor data. We make the following two hypotheses. First, we hypothesize that we can model the everyday behavior of a smart home resident using smart home sensor data. Second, we also hypothesize that there exists a relationship between everyday behavior and cognitive/physical health and that the learning algorithms can learn these relationships and use them to predict

the cognitive and physical health.

Based on these two hypotheses, we ask the following four research questions.

1. How can we characterize everyday behavior of a smart home resident using data from smart home sensors and learning algorithms such that the characterizations correlate with the clinical measures of everyday behavior using cross-sectional and longitudinal smart home sensor data?

2. How can we predict cognitive or physical health using such sensor-based characterization of everyday behavior?

3. How can we model the daily routine of a smart home resident using longitudinal sensor data?

4. How can we detect changes in the daily routine of a smart home resident?

Researchers have demonstrated methods to characterize simple scripted activities using data from smart home sensors [34, 63]. However, characterizing real-life activities using sensors has not previously been addressed. This is an important question because real life activities are often complex, and their characterizations are essential to understanding an individual's everyday behavior. We will answer this question first using cross-sectional data in Chapters 4 and 5 and using longitudinal data in Chapters 6 and 7.

First, we utilize the sensor data from the cross-sectional study to model the quality of simple and complex real life-like activities. This is an important research question because it allows us to validate the relationship between the sensor-based characterizations of activity quality and clinicians-measured activity quality measurements in a controlled experimental setting. In such studies, we can ask participants to perform a given set of activities and clinicians can "rate" the quality of the activities by directly observing them.

Extending the relationship between sensor-based measurements and clinical measurements obtained from the cross-sectional setting to a similar relationship in longitudinal setting is an important research problem because the actual smart home sensor data collected by monitoring an individual over a period of time will be longitudinal in nature. In addition, the changes in everyday behavior of an individual are more likely to be reflected in longitudinal data than in cross-sectional data [137]. We validate our approach by correlating longitudinal sensor-based measures of everyday behavior with longitudinal clinical measures of everyday behavior.

We explore how we can use learning algorithms to learn the differences in patterns of everyday functioning to predict cognitive and physical health scores. This is an important research question because it is the first step toward developing in-home intelligent systems that help clinicians and caregivers by predicting the health and well-being of an individual. While we lay out the framework of the cognitive

assessment agent in chapter 3, we answer this research question by providing evidence that differences in patterns in everyday day behavior can be used to predict cognitive health in Chapters 4, 5, 6, and 7. We answer the first two questions using both cross-sectional and longitudinal data.

We propose to design algorithms to detect changes in a person's everyday behavior by collecting sensor data of individuals longitudinally over a period. This is an important research question because it allows clinicians and caregivers to understand the actual daily routine of smart home resident and allows researchers to study the relationship between course of decline in completing activities and decline in cognitive health over a period. We limit the scope of the current research to the following three topics.

1. We focus on modeling measures of everyday behavior, especially activities of daily living. We exclude other measures such as gait parameters and other physiological signals such as heartbeat.

2. We only use data collected from unobtrusive environmental sensors to characterize measures of everyday behavior.

3. In an unconstrained setting, we develop algorithms for homes with single residents. The algorithms can be extended to multiple residents by tracking and identifying multiple residents based on the research effort by researchers includ-

ing Crandall [36].

In this work, we offer several contributions to the field. First, we present cognitive and physical assessment systems as intelligent agents, in which agents perceive everyday behavior using sensors and use this information to make predictions regarding the cognitive and physical health of a smart home resident.

Second, we introduce the design of a smart home-based cross-sectional experimental study in which we bring a large pool of participants into the smart home to perform different real life-like activities. We extend this cross-sectional study to a longitudinal study in which we monitor multiple individuals over two years in their own home using smart home sensors.

Using the sensor data collected from these two studies, we propose machine learning algorithms to characterize everyday activities. First, we propose algorithms to model the set of simple and complex activities when participants performed these activities in the smart home apartment. We utilize the information of how other people in the population performed the same activities to predict their cognitive health. Next, we propose algorithms to model everyday behavior of a smart home resident when the resident performs activities in their own home environment. We use automatically recognized activities using smart home sensors to characterize their everyday behavior. In both of these studies, we demonstrate that such characterizations are clinically relevant by correlating them with standard clinical measurements

of everyday behavior and provide evidence that they can be used for predicting cognitive/physical health scores. Finally, we propose an activity curve to model the daily activity routines of a smart home resident and demonstrate how activity curves can be used to detect changes in daily routine. We use a smart home resident's past behavior as a baseline to better understand their current cognitive and physical health.

CHAPTER 2. BACKGROUND

All of the smart home testbed described in this dissertation make use of the CASAS Technology Platform (CTP). CTP is a smart home implementation designed to sense the home environment. It uses motion, door, and object sensors to sense the space, middleware to collect and record the sensor events, and the database to store the sensor events [36].

2.1 Smart home testbed

2.1.1 CASAS Smart home testbed

The WSU CASAS[1] team has deployed multiple CASAS smart home testbeds. The testbeds are used to conduct various studies related to assisted living. Some examples of these studies are detection and recognition of Activities of Daily Living [138], tracking residents [36], studying energy usage [25], and prompting individuals to perform needed activities [39]. Below, we first describe the smart home testbeds, including the sensor technologies that are used in those testbeds and the middleware

[1] casas.wsu.edu

that collects and stores sensor data on in a relational database.

2.1.2 The Kyoto smart home testbed

The Kyoto smart home testbed is an on-campus testbed located at Washington State University. This smart home testbed is an apartment that has a living room, a dining area, and a kitchen on the first floor, and two bedrooms, an office, and a bathroom on the second floor. The apartment is instrumented with motion sensors on the ceiling, door sensors on cabinets and doors, and item sensors on selected kitchen items. The testbed also has temperature sensors in each room, sensors to monitor water and burner use, and a power meter to measure electricity consumption. Item sensors are placed on different items in the apartment to monitor their usage. Figure 2.1 shows the sensor layout in the CASAS smart home testbed. The experiments are performed in the first floor of the apartment while the experimenter monitors each participant upstairs via a web camera and remotely communicates with the participant using an intercom system.

Sensor events are generated and stored while participants perform the activities. Each sensor events is represented by four fields: date, time, sensor identifier, and sensor message. By examining data files and testbed floor plans, our team members annotate the sensor events with the activity that was being performed when the sensor

event was generated. A sample of collected sensor events and their corresponding labeled activities are shown in Figure 2.2. The CASAS middleware collects sensor events and stores the data in an SQL database. All software runs locally using a small Dream Plug computer.

The Kyoto testbed was used to perform two sets of tasks: Eight Activities and Day Out task. Participants in our cross-sectional study first completed the Eight Activities tasks and then completed the complex Day Out Task. These two set of tasks, the related studies, and their corresponding results are explained in Chapters 4 and 5

2.1.3 CASAS longitudinal smart home testbeds

The CASAS smart home testbeds that are used for the longitudinal study are actual single-resident apartments, each with at least one bedroom, a kitchen, a dining area, and at least one bathroom. The sizes and layouts of these apartments vary between homes. The homes are equipped with combination motion/light sensors on the ceilings and combination door/temperature sensors on cabinets and doors. These sensors in the smart home testbeds unobtrusively and continuously monitor the daily activities of its residents. The CASAS middleware collects these sensor events and stores the data on a database server. Appendix C shows the layout of all

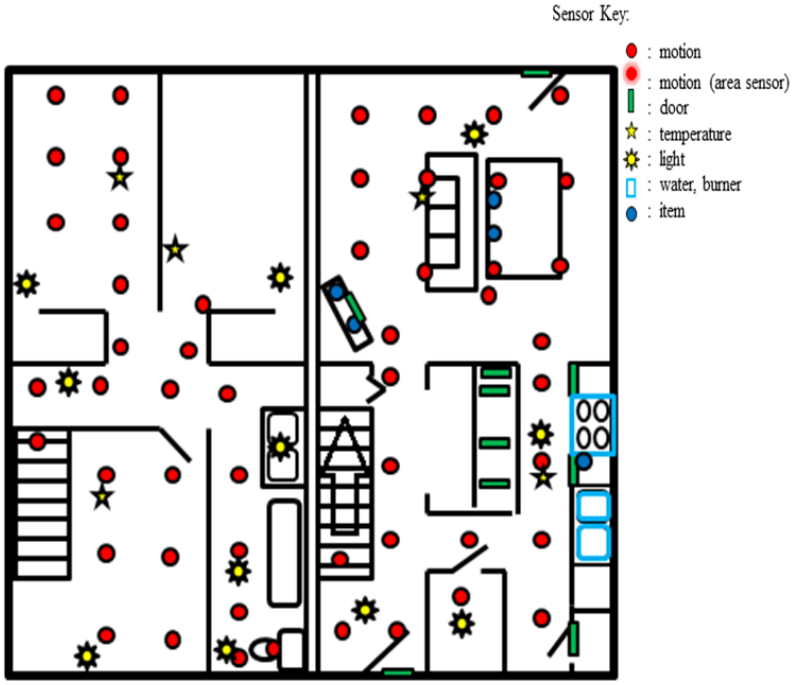


Figure 2.1: The Kyoto smart home testbed floor plan and sensor layout.

18 longitudinal smart home testbeds and the sensor placements in those apartments [29].

The residents perform their normal activities in these smart apartments, unobstructed by the smart home instrumentation. CASAS middleware that runs locally and is installed in the apartments collects the sensors events and uploads the data to the main database server. The format of the sensor events from the longitudinal smart home testbed is similar to those generated from the Kyoto testbed as shown in Figure 2.2.

Date	**Time**	**Id**	**Message**	**Activity**
2009-07-11	09:00:41.05	M042	ON	Bus/Map-end
2009-07-11	09:00:41.09	M025	OFF	Change-start
2009-07-11	09:00:42.05	M002	ON	
2009-07-11	09:00:42.82	T008	23	
2009-07-11	09:00:42.82	P001	23	
2009-07-11	09:03:58.82	M012	OFF	
2009-07-11	09:04:42.82	D001	OPEN	Magazine-end
2009-07-11	09:05:02.82	M009	ON	Change-end

Figure 2.2: Sensor file format from the Kyoto smart home testbed and sample annotations. Sensor IDs starting with M are motion sensors, D are door sensors, T are temperature sensors, and P are power usage sensors. The data is annotated with the start and ends of the subtasks.

2.1.4 Smart home sensors

The smart home testbeds that we describe here use almost identical sensor setup and middleware design to the Kyoto testbed that was described earlier. The sensors continuously generate sensor events when they monitor human activities. The middleware collects and records the sensor events in the database. We explain the three major types of sensors that these testbeds use in the following section.

Passive infrared motion detector

A PIR motion detector (Figure 2.3) detects the presence or absence of motion. Two types of motion sensors, area sensors and downward facing sensors, are used in the CASAS smart home testbeds. The area sensors are installed in a room such that their field of view is limited to a single room. While they can signal when an occupant enters the room, they cannot signal identification and location of the occupants in the room. In contrast, the downward facing sensors are mounted on the ceiling with the lens facing downwards such that their field of view is limited to directly below them. This arrangement allows achieving a more focused view of the space. To cover the maximum possible space, multiple downward facing motion detectors are placed in the rooms. Therefore, when they are triggered, they can provide information about the occupant's location. They send ON and OFF events to the middleware. An ON message indicates that motion was sensed in the sensor field of view. An OFF

Figure 2.3: Passive infrared motion detector used on CASAS smart home testbeds (adapted from [36]).

message indicates that sensed motion ceased in that field of view. Figure 2.3 shows the passive infrared motion sensors.

Magnetic door sensors

Magnetic door sensors are installed on bedroom, kitchen cabinet, and refrigerator doors to detect a door "OPEN" or "CLOSE" event. Magnetic door sensors are magnet driven reed switches that send "OPEN" events to the middleware when the magnet moves away from the reed switch. Similarly, a door sensor sends a "CLOSE" event when the magnet moves back into place. Figure 2.4 illustrates examples of magnetic door sensor usage in the Kyoto smart home testbed.

Figure 2.4: Magnetic door sensors on the door (left) and cupboard (right) (adapted from [36]).

Item sensor

Item sensors (Figure 2.5) detect presence or absence of important items such as medicine dispensers and cookware in the home. They were designed using contact switches and plates. When an item is removed from the plate, the switch is depressed which sends an "ABSENT" event to the middleware. Similarly, it sends a "PRESENT" event when the object is placed on the plate.

2.1.5 CASAS middleware

The CASAS middleware software architecture components are shown in Figure 2.6. Control flows up from physical components through the middleware to software applications when a signal is received. Similarly, control moves down from the

Figure 2.5: Item sensor(top) and use of item sensors in kitchen cupboard in smart home testbed (adapted from [36]).

application layer to the physical components while taking actions [29].

The CASAS physical layer contains hardware components including sensors and actuators. Examples of these sensors are motion sensors, item sensors, and object sensors. The architecture utilizes a ZigBee wireless mesh, which communicates directly with the hardware components.

The CASAS middleware provides services and information flow between various software and hardware components comprising the smart home. The middleware layer is governed by a manager. The manager provides named broadcast channels that allow component bridges to publish and receive messages. When a raw event is published, the manager assigns the event an id and adds a time stamp. Similarly,

other channels are used to publish or receive messages corresponding to actuator control commands or to identify entities (e.g., people, pets, other movable items) at the site.

Every hardware and software component of the CASAS Smart Home in a Box (SHiB) architecture communicates via a customized XMPP bridge to the manager. Examples of such bridges are the ZigBee bridge, the Scribe bridge which records sensor readings in a relational database, and bridges for various applications such as visualization, learning, and decision making. The Zigbee agent encompasses mechanisms that are necessary to connect and manage the ZigBee network. The CASAS smart home components use a Zigbee network to communicate by publishing and subscribing messages sent along various channels. Similarly, the Scribe agent archives messages in permanent storage. The data is organized and stored in a PostgreSQL database [29, 36].

2.1.6 Activity recognition algorithm

Activity recognition algorithms label activities based on readings (or events) that are collected from smart environment sensors. The challenge of activity recognition is to map a sequence of sensor events onto a value from a set of predefined activity labels. These activities may consist of simple ambulatory motion, such as

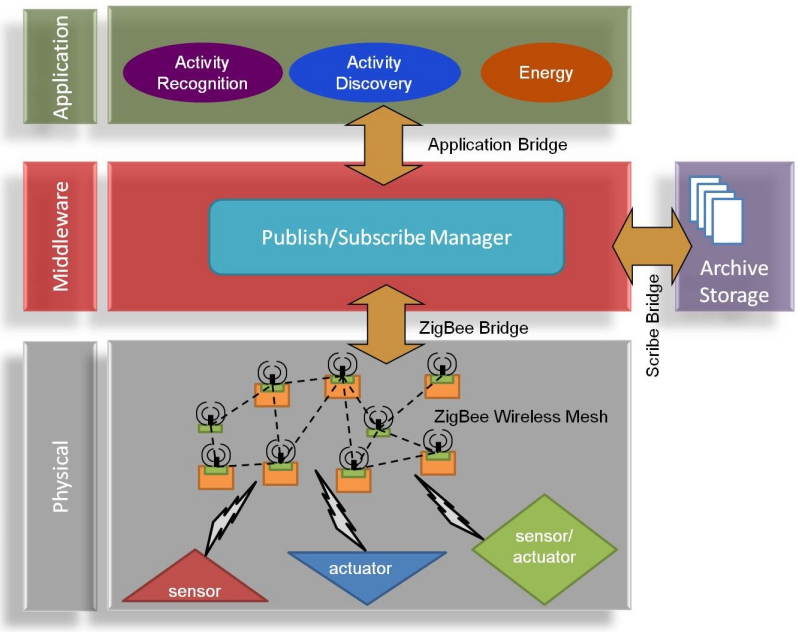


Figure 2.6: CASAS software middleware architecture.

walking and sitting, or complex basic or instrumental activities of daily living, depending upon what type of underlying sensor technologies and learning algorithms are used. For example, to recognize posture [93], gesture [79], or ambulatory activities such as walking, running, and sitting, researchers use data from wearable sensors such as accelerometers that are attached to the body [91]. Similarly, researchers use data from environmental sensors such as infrared motion sensors and pressure mats to recognize complex activities such as sleeping, eating, and cooking [138]. For activities that involve interacting with objects such as washing hands, researchers use data from RFID tags or shake sensors [19, 108, 116].

A number of survey articles provide an overview of method for activity recognition [3, 20, 26, 125, 150, 151] on simulated and real-world datasets. For example, Cook [31] explore activity recognition using Naive Bayes classifier and smart home sensor data and Bao et al. [85] and Ravi et al. [124] for accelerometer data. Hidden markov models [7, 83, 138, 158] and conditional random fields [16, 31, 76] have also widely been used or activity recognition. Other researchers have used decision trees [85] and discriminative learning algorithm such as support vector machines [21]. These underlying learning algorithms have been combined using boosting and other ensemble methods to create a robust activity recognition algorithm [84, 123, 156]. To evaluate the performance of the activity recognition algorithm, Ward et al.[158] and Bulling et al. [21] introduce a number of time-based and event-based performance metrics and Reiss etal.[125] discuss the effectiveness of hold-one-activity-out evaluation of activity recognition algorithms. Approaches to the activity spotting using data from wearable sensors is described in Ogris et al. [102] and Amft [8].

2.1.7 AR

Our activity recognition algorithm, AR [33, 78], recognizes activities of daily living, such as cooking, eating, and sleeping using streaming sensor data from environmental sensors such as motion sensors and door sensors. These motion and door

sensors are discrete-event sensors with binary states (On/Off, Open/Closed). Human annotators label one month of sensor data from each smart home with predefined activity labels to provide the ground truth activity labels for training and evaluating the algorithm. The inter-annotator reliability (Cohen's Kappa) values of the labeled activities in the sensor data ranged from 0.70 to 0.92, which is considered moderate to substantial reliability. We use the trained model to generate activity labels for all of the unlabeled sensor data.

AR identifies activity labels in real time as sensor event sequences are observed. We accomplish this by moving a sliding window over the data and using the sensor events within the window to provide a context for labeling the most recent event in the window. The window size is dynamically calculated based on the current sensor. Each event within the window is weighted based on its time offset and mutual information value relative to the last event in the window. This allows the events to be discarded that are likely due to other activities being performed in an interwoven or parallel manner. We calculate a feature vector using accumulated sensor events in a window from the labeled sensor data collected over a month. The feature vector contains information such as time of the first and last sensor events, temporal span of the window, and influences of all other sensors on the sensor generating the most recent event based on mutual information. Currently, AR recognizes the activities we monitor in this project with 95% accuracy based on 3-fold cross validation. An

2009-05-15	09:00:41.05	M042	ON		2009-05-15	09:00:41.05	M042	ON	Cook
2009-05-15	09:00:41.09	M025	OFF		2009-05-15	09:00:41.09	M025	OFF	Eat
2009-05-15	09:00:42.05	M002	ON	$\longrightarrow$	2009-05-15	09:00:42.05	M002	ON	Relax
2009-05-15	09:00:42.82	M028	ON		2009-05-15	09:00:42.82	M028	ON	Cook
2009-05-15	09:00:43.16	M027	ON		2009-05-15	09:00:42.82	M028	ON	Cook

Figure 2.7: Activity annotated sensor data. Sensors IDs starting with M are motion/light sensors and IDs starting with D are door/temperature sensors. The sensor events on the left describe the individual sensor readings while the sensor events on the right are annotated with activity labels.

example of activity-labeled sensor data is presented in Figure 2.7 [33, 78]. More details on this and other approaches to activity recognition are found in the literature [32]. We use the AR activity recognition algorithm to label individual sensor events with corresponding activity labels in sensor data collected from the longitudinal smart home studies.

2.2 Clinical screening

Participants for all of the studies used in this dissertation underwent comprehensive clinical tests in a laboratory setting. As detailed in Tables 2.1 and 2.2, the

initial screening procedure for the middle age and older adult participants consisted of a medical interview, the clinical dementia rating (CDR) instrument [97], and the telephone interview of cognitive status (TICS) [18].

Interview, testing and collateral medical information (results of laboratory and brain imaging data when available) were carefully evaluated to determine whether participants met clinical criteria for MCI or dementia. Inclusion criteria for MCI (see Table 2.1) were consistent with the criteria outlined by the National Institute on Aging-Alzheimer's Association work group [5] and the diagnostic criteria defined by Petersen and colleagues [114, 115]. The participants met criteria for amnestic MCI, non-amnestic, as determined by scores falling at least 1.5 standard deviations below age-matched (and education when available) norms on at least one memory measure (see Table 2.1). Participants with both single-domain and multi-domain MCI (attention and speeded processing, memory, language, and/or executive functioning) are represented in this sample. Participants in the dementia group met diagnostic and statistical manual of mental disorders (DSM-IV-TR) criteria for dementia [10] and scored 0.5 or higher on the clinical dementia rating instrument. The TICS scores for individuals with dementia ranged from 18 to 29 (M= 24, std= 3.71).

2.2.1 Clinical tests

Additionally, clinicians administered standardized clinical, cognitive, and motor tests every six months to residents of the longitudinal smart home testbeds. As detailed in Table 2.3, these tests included Timed Up and Go Test (TUG) and a global measure of cognitive status (RBANS). The administered clinical tests are standardized and validated measures that provide indication of mobility-based health and cognitive health. Table 2.3 provides a brief description of each clinical test and the measure that was employed from each test.

Table 2.1: Inclusionary and exclusionary criteria for the MCI group.

Inclusion criteria for MCI group:

1. Self-report or knowledgeable informant report of subjective memory impairment for at least 6 months, as assessed by direct questioning during initial screening interview.

2. Objective evidence of impairment in single or multiple cognitive domains (memory, executive, speeded processing, and/or language), with scores falling at least 1.5 standard deviations below age-matched (and education when available) norms. Test listed by domain with reference to norms used in parentheses:

    - Memory: Memory Assessment Scale list learning and long-delayed free recall [161]; Brief Visual Memory Test learning and long-delayed free recall [14]
    - Executive: Delis-Kaplan Executive Functioning Scale total correct from the Letter Fluency and Design Fluency subtests [42] ; Trail Making Test, Part B total time [67];Wechsler Adult Intelligence Scale-Third Edition Letter-Number Sequencing subtest total correct [160]
    - Speeded processing: Trail Making Test, Part A total time [67]; Symbol Digit Modalities Test total correct written and oral [139]
    - Language: Boston Naming Test total correct [160] ; Delis-Kaplan Executive Functioning Scale Category Fluency subtest total correct [42]

3. Preserved general cognitive functions as confirmed by a score of 27 or above on the TICS (equivalent to the normality cutoff score of 24 on the Mini Mental Status Exam) [95] .

4. No significant impact of cognitive deficits on the participant's daily activities, as confirmed by a total CDR score of no greater than 0.5, which is consistent with minimal change in the participant's habits.

5. Nonfulfillment of the DSMIV-TR criteria for dementia (American Psychiatric Association, 2000), confirmed by reviewing screening data, neuropsychological testing data, and any available medical records.

Table 2.2: Inclusionary and exclusionary criteria for the healthy older adult and dementia groups.

Initial screening for all middle age and older adult participants:
1. Medical interview to rule out exclusionary criteria of history of brain surgery, head trauma with permanent brain lesion, current or recent (past year) self-reported history of alcohol or drug abuse, stroke, or a known medical, neurological or psychiatric cause of cognitive dysfunction (e.g., epilepsy, schizophrenia).
2. Clinical dementia rating instrument to assess dementia staging [163] .
3. Telephone interview of cognitive status [18] to assess cognitive status and exclude significantly impaired participants who would be unable to complete the assessment.
Inclusion criteria for healthy older adult controls:
1. Reported no history of cognitive changes.
2. Scored within normal limits on the TICS.
3. Scored a 0 on the clinical dementia rating.
Inclusion criteria for dementia patients:
1. Met DSM IV-TR criteria for dementia [10]
2. Scored 0.5 or higher on the clinical dementia rating.

Table 2.3: Variables in our standard longitudinal clinical dataset.

Variable name	Description
Repeatable Battery for the Assessment of Neuropsychological Status (RBANS)	RBANS [121]. This global measure of cognitive status identifies and characterizes cognitive decline in older adults.
Timed Up and Go (TUG)	TUG [117]. This test measures basic mobility skills. Participants are tasked with rising from a chair, walking 10 feet, turning around, walking back to the chair, and sitting down. The TUG measure represents the time required for participants to complete the task at a comfortable pace.

CHAPTER 3. MONITORING EVERYDAY ABILITIES AND COGNITIVE HEALTH USING PERVASIVE TECHNOLOGIES: CURRENT STATE AND PROSPECT

3.1 Introduction

Recent advancements in pervasive sensor technologies and learning algorithms have made continuous and unobtrusive monitoring and analysis of human activities in the home environment a reality [30]. For example, sensors are embedded in a person's home environment and in everyday objects. These sensors often unobtrusively monitor and collect human behavior data. By analyzing and visualizing such sensor data with algorithms, we can understand a person's behavior. For example, we can determine their sleep pattern by analyzing when they go to sleep at night and when they wake up in the morning, or we can identify the typical timing for activities of daily living such as eating breakfast and washing dishes. Understanding everyday cognitive abilities and performing cognitive assessments are two major uses for this sensor data.

Sensors that continuously collect everyday data of an individual offer a wealth

of information about their everyday behavioral events. Analysis of sensor data can reveal patterns of daily activities that can be studied to identify regularly completed activities and activities that were completed with difficulty. Similarly, information about whether the individual is regularly taking their medication or is having difficulties with sleep can be found. Furthermore, sensor data collected over a time period provides insights on day-to-day changes and trends in activity patterns. The fact that smart homes collect information-rich behavior data under real life circumstances makes them invaluable tools to understand an individual's real life everyday abilities.

In this chapter, we review pervasive in-home technologies to monitor everyday behavior in the home environment and perform cognitive assessment using these technologies. In particular, our goal is to highlight current practices for in-home sensor technologies that monitor everyday behavior and perform cognitive assessment, and to identify future research directions of these technologies and present the challenges that lie ahead.

3.2 Cognitive assessment systems

Clinicians use cognitive assessment systems to assess the cognitive health of an individual. They diagnose the type and severity of cognitive difficulties. Clinicians can use outcomes of the assessment systems to recommend treatments and decide on the

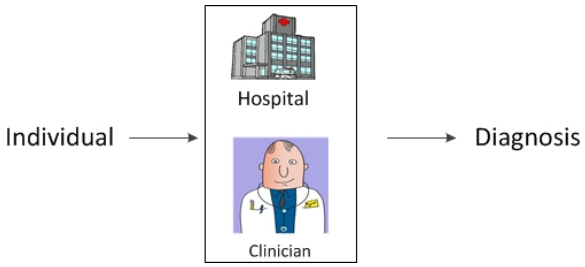


Figure 3.1: Traditional cognitive assessment systems.

required level of care and support. In addition, the treatments for cognitive decline are more effective when treated at early stages, before a cognitive disease causes irreversible damage to the brain. Thus, cognitive assessment systems can be used for early detection and management of cognitive decline. This slows the progression of the disease and consequently prolongs the individual's independent living and allows family members and caregivers more time to make appropriate decisions [165, 141].

In traditional cognitive assessment systems, conventional methodologies such as pen and paper-based tests are used to diagnose the type of cognitive difficulties [165]. Patients often visit the clinician who evaluates their cognitive health by administering various tests and using different scoring methods. Figure 3.1 illustrates this concept. Some examples of such standardized and validated tests are the Mini Mental State Examination [55] and the Modified Mini-Mental State Examination [148]. However, traditional clinical tests are administered in a laboratory. As a result, they require an individual to travel to clinics. They are often administered at an advanced stage

Table 3.1: Everyday functioning measures affected by cognitive health.

Categories	Everyday functioning measures affected by cognitive health
Computer usage	Keyboard and mouse usage, typing speed, performances in computer games
Mobility	Ability to move around, climb stairs, stride length
Gait	Gait velocity, gait balance,
Everyday functioning	Ability to initiate and complete activities of daily living such as bathing, toileting, eating etc.

when there are concerns of severe difficulties. In addition, they are administered infrequently since they have a long testing time and are expensive to administer. Since they are administered outside of the individual's home environment, i.e., in laboratories, they have limited ability to capture everyday difficulties of an individual. In another word, many of these tests are not considered to be ecologically valid [23]. These limitations in traditional settings have prompted researchers to pursue new directions of more ecologically valid methods of detecting difficulties in individuals in real life settings using pervasive computing solutions. Table 3.1 lists various examples of everyday functioning measures affected by cognitive health.

3.3 Cognitive health and everyday functioning

Everyday functioning in an individual is an important neuropsychological construct that clinicians try to understand. Several clinical studies support a relationship between daily behavior and cognitive or physical health [35, 132, 130]. While functional decline in at least one of the domains is a criterion for diagnosing dementia, individuals diagnosed with Mild Cognitive Impairment, a transition between normal cognition and dementia, also have difficulties completing complex activities of daily living such as managing finances. In addition, decline in everyday functioning is associated with reduced quality of life, risk of institutionalization, caregiver burden, and financial costs [142, 65]. The systematic characterization of everyday functioning and understanding the course of decline improve understanding of cognitive deficits that affect everyday abilities. This can pave the way for development of new methods to overcome the difficulties associated with impairments that can consequently prolong independent living [68]. Thus, early detection of decline in everyday functioning has important clinical and research applications that help clinicians understand patients' difficulties in everyday life and the decline in their cognitive health.

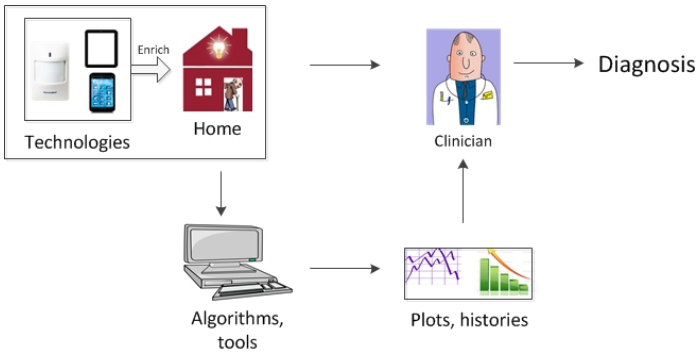
Table 3.1 lists various everyday functioning measures that are affected by decline in cognitive health. With decline in cognitive health, patterns of decline in these measures such as difficulties in completing activities independently and diminished

ability to move around can be observed.

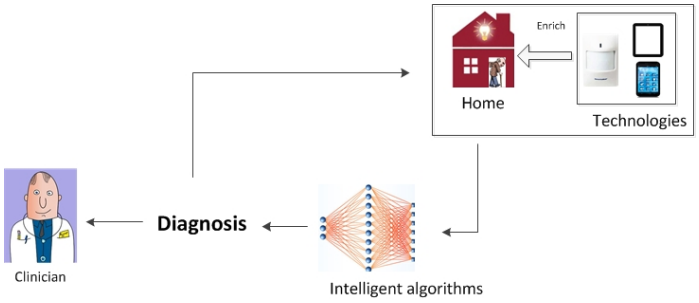
3.4 Technology and cognitive assessment system

Information source such as environmental, motion, and object sensors and video cameras can be installed in an individual's home. They can continuously collect everyday behavior data that is rich in information. For example, analyzing this data can provide detailed information about the individual's sleep pattern such as when and for how long the individual sleeps, and how many times he goes to the restroom during the night (bed to toilet transition). The insights that smart home technologies provide help researchers and clinicians better understand a person's everyday abilities and make informed decision about their cognitive health. In addition, electronic devices such as smart phones, tablets, and computers are technological gadgets that we continuously use in our daily routines. The embedded sensors including accelerometers, gyroscopes, and touch pads collect data when the individual interacts with these devices [70]. Analyzing the collected data reveals how the person behaves daily as well as fluctuations in routine behavior. These concepts are illustrated in Figure 3.2.

The information that technologies collect in real-life settings has motivated researchers to use pervasive in-home technologies to measure a person's everyday abilities and glean insights on their cognitive health. Formally, we define a *technology-*



(a) Technology-augmented agent assessment system



(b) Fully automated technology-assisted agent assessment system

Figure 3.2: Technology based cognitive assessment system.

assisted cognitive assessment system as a cognitive assessment system that uses technology to measure everyday functioning measures in-home and assists clinicians to make informed decision by providing in-depth real life data. Typically, in-home cognitive assessment systems use sensor technologies to collect the data. Intelligent algorithms analyze this data to provide valuable insights.

We can view a technology-assisted cognitive system as an *intelligent agent* in

which modern technologies enhance the traditional testing environment providing new perceptions to the learning agent, clinician, or intelligent learning algorithm, as illustrated in Fig 3.3. [30, 128]. Table 3.2 lists various components of the agent. Based on the role of technology, we classify them into three different categories:

1. In a *traditional agent assessment system*, clinicians are the agents who administer clinical tests and collect information. They perceive this information, reason and interpret them using their own expertise, and take actions such as conducting diagnosis and planning treatments.

2. In a *technology-augmented agent assessment system*, sensors are added to the traditional clinical environment. Intelligent algorithms provide their own perception of the individual's task performance based on sensor data, which expands the set of information passed to the reasoning/learning agent. Clinicians or caregivers act as learning agents. Fig. 3.2a illustrates this concept.

3. In a *fully automated technology-assisted agent assessment system*, the agent perceives information from its environment (sensors) and uses this knowledge to take actions (perform assessment) without human assistance. For example, these agents perceive sensor information related to an individual's everyday behavior in their home environment and alert the residents or caregivers when they detect significant deviations in everyday behavior. Fig. 3.2b illustrates

this concept.

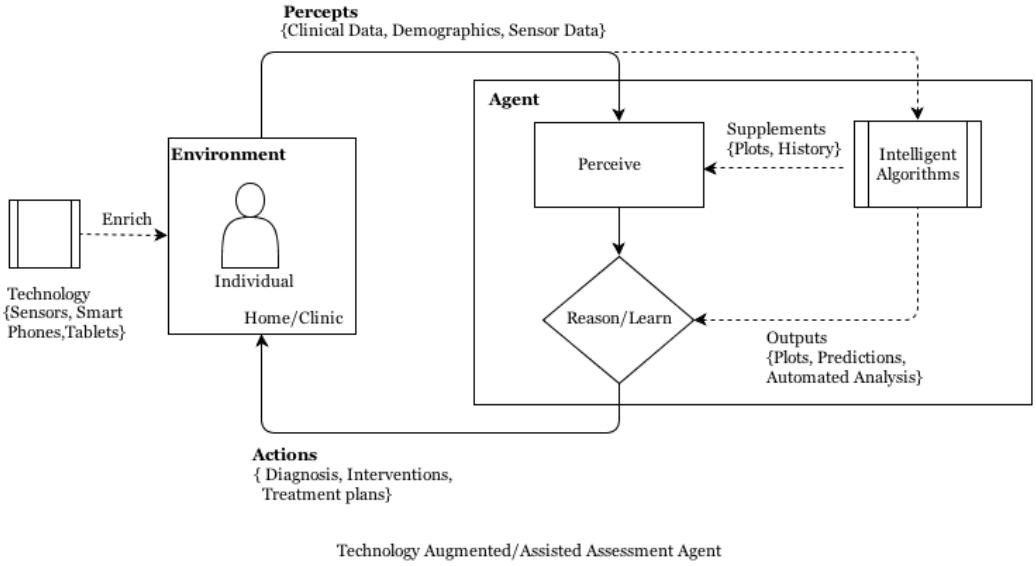


Figure 3.3: Technology based assessment agent.

3.5 Technology-assisted cognitive assessment

3.5.1 Cognitive assessment using computing devices

Cognitive assessment using computing devices can be an alternative to the traditional pen and pencil based tests. Clinicians can perform cognitive assessments by (i) utilizing data collected from computerized cognitive test suites that a participant completes on those devices, and (ii) by monitoring the personal cognitive digital behavior of a participant. These systems utilize sensor technologies present in the device

Table 3.2: Elements of technology assisted assessment agent.

	Traditional agent	Technology augmented agent	Automated agent
Agents	Clinicians, nurses	Clinicians, nurses	Learning algorithms
Perception	Standardized tests, Care giver questionnaires	Standardized tests, Sensor data, Tablet and smart phone data	Sensor data related to daily behavior, Tablet usage data
Actions	Diagnosis, medication, level of care	diagnosis, medication, Level of care	Interventions, diagnosis, alerts
Environment	Clinical labs	Traditional environment enriched with technology	Home

and make assessment process flexible by making it accessible at home. Individuals can complete them online and multiple times on any supported platform with limited cost and burden.

Compared with traditional testing systems, computerized cognitive assessment testing systems offer several advantages [99]. In contrast to traditional paper-based interface, computerized test systems can leverage digital interface to present end users with interactive and dynamic contents. The test systems can be customized according to end users requirements or testing methodologies. For example, tests can be adapted based on how a user responds to test questions, or they can be geared toward testing a particular cognitive difficulty. Furthermore, algorithms can minimize the errors caused by human biases and judgments by automating various steps in the tests, including scoring, and administration. Measurement constructs such as reaction time can be accurately measured while the system collects additional test metadata such as time spent in each component of the tests and the total number of pauses and breaks taken [99]. Overall, the computerized cognitive assessment tests offer the flexibility of taking the tests at home, enhancing the traditional testing environment with resources, and supplementing measurements with metadata. Therefore, they are in-home alternatives to traditional clinic-based pen and paper tests.

The other class of assessment systems that uses computing platforms performs non-traditional assessment based on an individual's cognitive digital behavior. The

cognitive digital behavior data are everyday computer usage behavior characterized according to how the individual interacts with the computer. Examples of these behaviors may include the person's keyboard or mouse usage patterns and their playing games on tablets and computers. We explain both of these methods next.

Cognitive assessment systems can use computers to perform cognitive assessments. Generally, the traditional pen and paper tests are adapted to computers. For example, the Finger-tapping test, the Stroop Interference test, and the CERAD Word List Learning Test are implemented as part of the computerized assessment battery. However, these tests have different sensitivity and specificity in classifying cognitive health of the participants. They also have different completion times. Some of these tests can be completed online while others require trained clinicians to be present [99, 141].

Cognitive assessment systems can utilize mobile platforms such as tablets and smart phones. Researchers have developed both tablet versions of the traditional pen and paper based tests and completely new test suites on mobile computing platforms. Using these portable devices, users can complete tests at any place and at any time. For example, users can complete the CANTAB test suite that includes tests for attention, visual memory, and executive functioning on their mobile devices. Often, these tests provide information that traditional pen and paper tests cannot. For example, in addition to the information about the correctness of the clock elements, the

tablet version of a clock drawing test provides "real-time pen interaction" data such as pausing tendency, pen pressure exerting patterns, and drawing sequences [77].

In addition, these devices are equipped with sensor technologies that provide additional methods of input. Possibilities include sensor-based inputs such as stylus, digital pen, and digital paper. One can exploit different sensors in the phone to develop a generic assessment suite or test suites that measure specific constructs. For example, Fouchenette [56] developed an iPod application to test long term memory. In this application, users can either use the touch screen to tap the correct answer, or type on a virtual keyboard, or use the microphone on the iPod to record the answer. These devices offer a sensor rich platform that provides researchers with an opportunity to develop tests that target specific cognitive domains while providing accurate measurements of the underlying constructs.

Virtual reality-based assessment systems represent another category of cognitive assessment systems that use computing devices. In virtual reality-based methods, participants wear special devices to manipulate entities in a virtual environment. Participants complete everyday activities in the virtual environment or complete the virtual reality based cognitive tests [74, 166]. A comprehensive review of virtual reality-based assessment systems are out of the scope of this work as they require specialized equipment and are not readily available at home.

Similarly, another variant of computerized cognitive assessment systems uses

cognitive digital behavior data. The data is collected during the time individuals use computers in their home environment playing computer games or using their keyboard and mouse. Assessments can be performed by deriving *computer cognitive metrics*, which serve as proxy measures to standardized neuropsychological tests [73]. For example, the Finger Tap Test measures motor control. In this test, participants repeatedly press a switch at a given time. The Finger Tap test can be simulated based on how individuals use keyboards at home. Typing speed during login can be used as a consistent and reliable proxy for measuring motor speed [71]. It was reported that individuals with cognitive difficulties have lower average scores and higher variability in typing speed compared with healthy individuals. Similarly, mouse and keyboard usage data can serve as a proxy for motor speed measurements. Cognitive digital behavior data collected over a period of time can be used to identify cognitive health trends, day-to-day variability, and comparing within subjects measured across a period of time [71].

Cognitive computer games are another approach to computerized assessment of cognitive health. Cognitive computer games are computer games with embedded algorithms to infer cognitive processes. Often, popular computer games such as FreeCell and Solitaire are modified to create cognitive computer games. When users play those games, the algorithms calculate different cognitive metrics that possibly correlate with measures obtained from traditional clinical measures. For example, Jimison

et al. [71, 72, 73] designed word games to obtain proxy measures correlating with the measures obtained from verbal fluency test. A verbal fluency test measures a person's ability to recall from long-term memory and their semantic processing ability by asking individuals to produce as many words as possible. To simulate this test, a computer game presents user with letters of scrambled words and asks them to make as many words as possible with incentives on longer words. The authors argue that game metrics such as word complexity (a mix of word length and frequency of use in the English language), the total number of words created, and the speed of generation of words approximate the standard verbal fluency measures. Using a similar concept, one can approximate other cognitive parameters such as memory, planning, and divided attention [70]. Such approximations can classify cognitively healthy subjects and those with mild cognitive impairment [72]. Cognitive computer games have also been implemented in mobile devices to assess the cognitive health of astronauts, who operate in a complex environment.

We categorize computer assessment systems as intelligent agents. The sensor technologies such as touch displays, mouse, stylus, and keyboards enhance (or substitute) the traditional pen and paper environment with several added capabilities and expand the perception of the agent by collecting diverse information from the environment. The clinicians are the agents who take necessary actions after interpreting their perception (the collected data) using their own experience. Alternatively, the

intelligent algorithms can study the data history and provide real time feedbacks and recommendations to the user or the caregiver.

3.6 Cognitive assessment using smart home sensors

Smart home sensors continuously monitor and collect data related to everyday abilities, for example, completion of activities of daily living, circadian rhythms of daily behavior, and the ability to move around. Clinicians can use sensor data to better understand the course of decline in everyday abilities. Such understanding helps them to design treatment plans, develop measures to overcome difficulties, and prolong an individual's independent living. In addition, using the sensor data, one can develop a more ecologically valid assessment system using data collected from real life events and predict a person's ability to live independently.

We explain smart home-based assessment systems using the agent framework as shown in Fig. 3.2a. The environment in the framework is the participant's home enhanced with sensors. These sensors continuously monitor everyday behavior of the resident and provide data and precepts to the learning and reasoning agent. Agents act and reason by interpreting this data and recommend or administer treatments. The agents can be clinicians, caregivers, or intelligent algorithms. Smart home sensors can perform cognitive assessment by monitoring key features of daily routines such

as mobility and activities of daily living.

3.6.1 Monitoring mobility

Mobility is the ability of an individual to move around their home environment and the community, while completing activities of daily living and maintaining an active social life [159]. Mobility impairments limit an individual's ability to maintain independence and quality of life and are one of the predictors of institutionalization among older adults [65].

Several cross-sectional and longitudinal clinical studies have investigated the relationship among mobility, gait disorders, and cognitive decline. They have shown a relationship between gait speed and risk of lower 3MSE score [54], gait abnormalities, such as unsteadiness and frontal gait disorders, and non-Alzheimer dementia [152], and time to walk 30 ft. and the onset of cognitive impairment [88]. Similarly, other studies have investigated relationships among cognitive dysfunction, risk of disability, ability to walk a few feet and climb stairs, ability to get out of bed and chair, and upper extremity strength [54, 155]. These studies concluded that mobility and gait impairment are associated with cognitive decline and prediction of dementia, future risk of hospitalization [144] and disability [60], and loss of executive function [35]. These studies demonstrate the importance of mobility and their relationship with

cognitive health. However, the data in clinical studies are usually collected using traditional methods of self-report, informant report, and performance-based measures. In contrast, sensors continuously monitor in-home mobility for longer duration and provide frequent measurements. Therefore, they can be valuable in understanding mobility in real life [61].

The mobility of individuals can be monitored using wearable sensors such as accelerometers and gyroscopes [61] as well as pressure sensor-wired shoe insoles [162]. The possibilities to recognize mobility and gait parameters also include ambient sensors such as passive motion sensors [12], indoor sensor mats [96], laser scanners [57] and [143] video cameras. The measured parameters using PIR motion sensors have been validated with the standard systems. Hagler et al.[61] collected in-home walking speeds using PIR motion sensors and validated it using GAITRite system. Ambient sensors are preferable to wearable sensors since they are unobtrusive. However, installing ambient sensors requires prior knowledge of the layout of the room and may pose difficulties for multiple-resident homes [61, 72].

The initial work that monitored mobility and predicted possible cognitive decline using mobility parameters analyzed continuously monitored sensor data using Semi Markov Model to make estimations [109]. Other researchers predict decline in cognitive health based on the gait speed estimated from sensor data [11]. In the home environment gait speed is estimated based on the timing of sensor events arranged in

a straight line down a hallway or narrow corridor. Studies have shown evidence for a relationship between in-home gait speed and abrupt changes in health condition, as well as a relationship between weighted correlation estimates of the gait speed and cognitive health [13]. However, estimating gait velocity using this technique poses challenges in multiple-resident homes. As one solution, Austin et al. [11] proposed modeling gait speeds using Gaussian Mixture Models for multiple residents in a smart home. They show the effectiveness of their technique by correlating the results from sensor data with the standard clinical assessment data. Researchers have also studied the relationship between the longitudinal trajectories of walking speed and speed variability and the cognitive health using latent trajectory modeling technique [45].

These modeling techniques use passive infrared motion sensor systems and require explicit arrangements of sensor systems to measure walking speed. Alternatively, one can use a Kinect system to measure walking speed, stride length, and stride time. The measurements using Kinect sensors were validated using marker-based motion capture system [143]. Such measurements can be used to detect fall and early onset of functional decline.

3.6.2 Monitoring everyday functioning (Activities of daily living)

Everyday functioning is the functional ability of an individual to complete daily activities to live competently and independently. Functional abilities such as eating, maintaining hygiene, and using bathroom are called basic activities of daily living. They are cognitively less demanding but are fundamental to living. Similarly, instrumental activities of daily living (IADLs) are another set of functional abilities. These cognitively demanding activities include driving, using telephones, and managing finances. For independent living, a person needs to complete activities of daily living. Measuring and understanding difficulties in everyday functional abilities are therefore important parts of gerontology research.

Previously, clinical studies have shown that individuals diagnosed with cognitive difficulty have more difficulties in completing IADLs when compared with healthy controls [9, 53, 110]. With incidence of more severe cognitive problems such as AD, individuals have difficulty in both initiating and completing basic activities as compared with healthy and MCI controls. Often, with progression of cognitive difficulty, a pattern of decline in abilities to complete higher order functional abilities (IADL) followed by basic functional abilities is observed [59]. Such pattern indicates that detection of functional changes may help detect early stages of cognitive decline or identify individuals with greater risk of decline [110]. Predicting cognitive health

based on the performance of activities of daily living is an active research area in neuropsychology and clinical research [130, 131].

Activity recognition and discovery algorithms can be used to study a person's everyday functioning using sensor data. The raw sensor data does not contain activity labels, and therefore it requires annotation before it can be used further. Activity recognition and discovery algorithms can annotate sensor events with activity labels. While activity recognition [27] algorithms map a sequence of raw sensor readings to a label indicating the activity that is being performed, activity discovery [122] algorithms discover sensor patterns of frequently conducted activities but cannot provide specific activity labels. These algorithms can accommodate various sensor modalities including environmental sensors, wearable sensors, and object sensors. They can also recognize complex and interweaved activities of daily living [138]. The output from activity recognition and discovery algorithms can be used to develop activity-tracking algorithms to track an individual's daily behavior.

Activity tracking algorithms monitor activities in daily life. They flag abnormal changes in activity patterns after discovering frequently completed activities [122]. They may also use domain specific rules to detect changes in circadian rhythms of the daily activities [153]. These abnormalities and changes in circadian rhythms are hypothesized to indicate problems in cognitive health. The activity tracking algorithms are particularly useful to track an individual's activity over a long period

of time and detect decline in their everyday abilities.

In other works, researchers hypothesized that individuals with cognitive difficulties have significantly lower *activity quality* than healthy controls. This hypothesis was drawn assuming that individuals with cognitive difficulties commit more significant errors while completing activities. The focus of this research is on development of intelligent algorithms that can predict activity quality and study the relationship between cognitive health and quality of activities of daily living. The data is usually collected under a constrained laboratory environment while groups of people with different cognitive difficulties perform predefined set of activities.

In one of such works, Hodges et al. [63] hypothesized that patterns of errors made while completing daily activities correlate with type and severity of cognitive difficulty. The authors assumed coffee-making activity as a proxy to everyday functioning and monitored the object usage using wireless RFID sensors while individuals with TBI completed the activity. They found a correlation between extracted set of features that characterized the task performance based on the object usage data and neuropsychological assessments tests. Researchers have also used ambient sensors to derive the quality of completed activities. Cook & Schmitter-Edgecombe [34] developed a Markov model to assess the completeness of five different activities of daily living: telephone use, hand washing, meal preparation, eating and medication use, and cleaning. Their model detected certain types of step errors, time lags, and

missteps. Both of these studies were conducted while participants performed simple daily activities. Researchers have also developed algorithms to monitor a specific activity such as dressing [92].

Monitoring activities with sensors have shown that activity quality and cognitive health are related in complex and real life activities [40, 41]. The authors classify individuals into cognitively healthy and cognitively unhealthy using activity quality as an input to learning algorithms. They also show a fairly strong correlation between direct observational scores that were observed on a set of activities of daily living and the sensor measurements of the same activities by trained neuropsychologists. The latter indicates that by using sensor data, learning algorithms can predict activity quality of both simple and complex activities, and that such predictions correlate with activity quality measurements performed by trained clinicians.

All the above-mentioned studies present tracking everyday functioning using sensor technologies as a promising candidate to monitor cognitive health. The shortcoming of these studies is that they were conducted under a laboratory setting using a limited set of predefined activities. Future research in this area must be focused on investigating the relationship between everyday functioning and cognitive health in real and unconstrained environments.

3.7 Longitudinal monitoring

In contrast to the cross-sectional studies of everyday functioning in which researchers collect data from a population at a certain time point, longitudinal studies collect repeated observations of everyday functioning of individuals continuously over a long period. These studies are better suited to address the questions related to within-individual changes and inter-individual differences in changes, trends, and trajectories [137]. Using longitudinal data, one can detect changes in data by taking a person's past as a baseline. Thus, to detect early indications of cognitive decline based on continuous monitoring of everyday behavior, analysis of longitudinal everyday functioning data is ideal. However, longitudinal analyses require data collected over a long period and pose additional challenges such as missing data and dropouts.

Previously, clinical researchers have studied patterns of changes and their trajectories in everyday functioning measures and their relationship to cognitive health. For example, Artero et al. studied the relationship between the ability to perform everyday activities (functional abilities) and cognitive disorder in 368 participants for three years. They found that individuals with mild cognitive deficits have more difficulties completing everyday activities when compared with cognitively healthy individuals. In a similar work, Wadley et al. [154] examined trajectories of changes in everyday functioning of 2358 participants over a three-year period and found that

the rate of decline of everyday functioning is higher in individuals with cognitive difficulties (mild cognitive impairment). Similarly, Peres et al. [112] followed nearly 1000 healthy individuals and individuals with dementia for ten years and found that among those healthy individuals, the ones who developed dementia at a later time had worse performance in complex IADLs compared with the healthy controls that did not develop dementia. They conclude that changes in daily pattern may constitute early markers of decline in cognitive health. The relationship between changes in mobility patterns and cognitive difficulties has also been studied in O'Connor et al [100].

However, clinical studies often use self-report and informant report-based methods since behavior simulation and direct observation methods of data collection procedures are expensive and labor intensive to carry out repeatedly. In contrast to self-report scoring, smart home sensor systems can continuously monitor and collect measurements of everyday behavior of the residents in their home environment.

A limited number of studies have investigated longitudinal monitoring of everyday functioning parameters using sensor data and studied their relationships with cognitive health. While some researchers have developed techniques to visualize gait, sleep, activity densities, and circadian rhythm over a long period, other researchers have used statistical modeling techniques to model the relationship between longitudinal sensor data and cognitive health. We discuss both of these approaches next.

Visualization techniques represent sensor data in a way that clinicians and caregivers can comprehend and visually detect changes in activity patterns and activity rhythms of the patient. In one of such works, Wang et al. [157] used motion sensor data to plot an activity density map, which is a visualization plot that represents levels of activities with different colors. Using a dissimilarity metric among activity density maps, the authors demonstrated techniques to track changes both in daily activity patterns over time and in physical and cognitive health. In another work, Virone et al. [153] presented techniques to model and visualize daily circadian rhythm of the activities and their deviations. They calculated time spent in each room of the smart home and the number of motion sensor events triggered per room. Such visualization technique has also been developed to visualize deviations in activities of daily living. Similarly, Kanis et al. [75] developed techniques to visualize activities in order to detect early indications of diseases with feedback from medical experts. These works focus solely on the development of an effective visualization tool for long term data monitoring [167].

While visualization techniques are helpful tools for both caregivers and clinicians to derive quick conclusions, they neither model the statistical relationship between health events and sensor data, nor do they generalize the relationship across the population. Thus, other researchers have quantified the relationship between sensor data and standard clinical scores. In one such work, Paavilainen et al. [106] found

lower daytime and higher nocturnal activity levels in individuals with dementia compared with healthy individuals. They also found statistically significant correlations between self-assessment of sleep quality and daytime vigilance. The activity signal data was collected using IST Vivago Wrist care system. Similarly, Paavilainen et al. [107] study changes in circadian activity rhythm using the same technologies as clinical observations of health status of the subjects and concluded a relationship exists between the two.

Researchers have also investigated the direct relationship between sensor data and standard clinical scores. For example, while Robben and Krose [126, 127] found a correlation between data obtained from the motion sensors and standard clinical assessment. The Assessment of Motor and Process Skills (AMPS) scores, Dodges et al. [45] studied the relationship between longitudinal trajectories of walking speed and speed variability and cognitive health using latent trajectory modeling technique. Suzuki and Murase [146] have shown a relationship between MMSE scores, health, and in-house movements. All these studies provide valuable evidence that longitudinal monitoring data can be used to make inferences about cognitive health.

3.8 Discussion

The recent advancements in sensor technologies and learning algorithms have made continuous monitoring of daily human behavior in their home environment a reality. We presented cognitive assessment systems as intelligent agents and discussed in-home technologies to monitor an individual's cognitive health. We highlighted clinical findings that suggested a relationship between everyday behavior and cognitive health and discussed smart home-based approaches to monitor everyday functioning. Interdisciplinary research effort among clinicians, neuropsychologists, and engineers is required to move this field forward.

In particular, future work on cognitive assessment based on computing platforms can utilize mobile platforms. By exploiting this sensor-rich platform, scientists can develop test suites that can be completed in an individual's home environment and can accurately measure the underlying cognitive construct. One can also adopt novel platforms such as full body gaming systems (e.g., Nintendo Wii and Microsoft Kinect) to perform cognitive assessments. Recent research suggests that full body gaming systems are becoming widely accepted among older adults [2]. Previous researchers have studied the acceptance of these gaming systems and their applications for physical rehabilitation for older adults. Future research should address the question of whether or not the data from these gaming systems can be used for cognitive

assessment or to simulate standard clinical tests.

Similarly, monitoring everyday abilities with smart home sensors have several research possibilities given the importance of understanding everyday abilities of older adults and the immense amount of information that sensor data contains. The initial studies have shown potential for sensor technologies to monitor daily activities in constrained laboratory setting and predict their quality. However, how their methods extend to unconstrained settings remains unanswered. In addition, extending activity recognition algorithms, which are well studied problems in pervasive computing, to develop activity tracking algorithms to track changes in activity patterns for a period of time is an open research question that needs to be addressed.

Recently, with the availability of longitudinal data, researchers have focused on development of algorithms to analyze long-term sensor monitoring data and detect early indications of cognitive decline [46]. Modeling of everyday functioning parameters and types of statistical and learning models required for detection of early indications of cognitive decline is an active area of research. The outputs from the models are valuable to end-users, especially to clinicians and caregivers, as they allow them to better understand a person's behavior. In addition, the data from large-scale studies can be used to answer questions about the population behavior. For example, one can answer the questions of whether or not an observed trend for a dementia group can be generalized across the population, or if MCI and dementia groups exhibit similar

trends.

Cognitive decline is a gradual and slow process. Researchers require continuously monitored data over a long period of time to detect cognitive decline. Currently, there are very few openly available long-term behavior sample data that researchers can use to develop algorithms. Very few of them have instances of known incidence of cognitive decline. Such lack of publicly available data complicates algorithm development and is an ongoing challenge for the field.

Ideally, a cognitive decline detection algorithm performs trend detection on longitudinal behavioral data. Such algorithms would take an individual's history as a baseline and perform the analysis. This algorithm requires convergence of traditional time series and longitudinal data analysis techniques and machine learning algorithms. In addition to analyzing each individual separately, it is also desirable to generalize the trend to the overall population and to observe if a detected individual trend can be generalized to the population. Still, this poses a great challenge since it requires large datasets. Thus, we stress the necessity of publicly available clinical ground truth data along with long-term sensor monitoring data to advance and motivate researchers to develop algorithms for performing in-home cognitive assessments.

After analyzing the results obtained from sensor measurements, the final step is to verify that the obtained results *align* with the results obtained from clinical data that are accepted by the community. This step ensures that the results have indeed

captured some existing underlying trends on a validated standard clinical dataset (see Appendix A).

In the next two chapters, we will first discuss two cross-sectional studies: the Eight Activities study and the Day Out Task study. These cross-sectional studies will allow us to compare activity performance across populations and to automate activity quality categorization based on activity performance differences. We will introduce algorithms to predict the activity quality of these activities and to automatically categorize cognitive health using collected smart home sensor data and machine learning algorithms. After discussing the cross-sectional studies, we will discuss a longitudinal study in which smart home sensor data is collected for over two years in a real life setting. We will discuss algorithms to model everyday behavior of a smart home resident and detect changes in the everyday behavior. We will then use this model to predict the cognitive and physical health of a smart home resident.

CHAPTER 4. CROSS-SECTIONAL ANALYSIS OF EIGHT ACTIVITIES

In this chapter, we first discuss the design of the smart home-based cross-sectional study in which a pool of participants performed eight different activities in our smart home testbed. We then present a machine learning algorithm that automatically predicts the activity quality of these activities utilizing sensor data that is collected while participants perform activities in the smart home testbed. We also present algorithm to perform cognitive health assessment to classify participants as cognitively healthy, MCI, or dementia based on activity quality modeled from sensor data.

4.1 The smart home testbed

Data are collected and analyzed using the Washington State University CASAS on-campus smart home testbed, an apartment that contains a living room, a dining area, and a kitchen on the first floor and two bedrooms, an office, and a bathroom on the second floor. For more details on smart home test bed, refer to Section 2.1.

4.1.1 Smart home activities

During the experiment, each participant was introduced to the smart home testbed and guided through a preliminary task in order to familiarize the participant with the layout of the apartment (see Chapter 2). The participant was then asked to perform a sequence of eight activities. Instructions were given before each activity and no further instructions were given unless the participant explicitly asked for assistance while performing the activity. The eight activities are:

1. *Household chore*: Sweep the kitchen and dust the dining/living room using supplies from the kitchen closet.

2. *Medication management*: Retrieve medicine containers and a weekly medicine dispenser. Fill the dispenser with medicine from the containers according to specified directions.

3. *Financial management*: Complete a birthday card, write monetary check, and write an address on the envelope.

4. *General activity*: Retrieve the specified DVD from a stack and watch the news clip contained on the DVD.

5. Household Chore: Retrieve the watering can from the closet and water all of the plants in the apartment.

6. *Telephone use/conversation*: Answer the phone and respond to questions about the news clip that was watched previously.

7. *Meal preparation*: Cook a cup of soup using the kitchen microwave and following the package directions.

8. *Everyday planning*: Select clothes appropriate for an interview from a closet full of clothes.

These activities represent instrumental activities of daily living (IADLs) [64] that can be disrupted in MCI, and are more significantly disrupted in AD. As there is currently no gold standard for measuring IADLs, the IADL activities were chosen by systematically reviewing the literature to identify IADLs that can help discriminate healthy aging from MCI [111, 149]. All IADL domains evaluated in this study rely on cognitive processes and are commonly assessed by IADL questionnaires [81] and by performance-based measures of everyday competency [44, 163]. Successful completion of IADLs requires intact cognitive abilities, such as memory and executive functions. Researchers have shown that declining ability to perform IADLs is related to decline in cognitive abilities [55].

In this chapter, we examine whether sensor-based behavioral data can correlate with the functional health of an individual. Specifically, we hypothesize that an individual without cognitive difficulties will complete our selected IADLs differently

than an individual with cognitive impairment. We further postulate that sensor information can capture these differences in quality of activities of daily living and machine learning algorithms can identify a mapping from sensor-based features to cognitive health classifications.

4.1.2 Experimental setup

Participants for this study completed a three hour battery of standardized and experimental neuropsychological tests in a laboratory setting, followed approximately one week later by completion of everyday activities in the smart home. The participant pool includes 263 individuals (191 females and 72 males), with 50 participants under 45 years of age (YoungYoung), 34 participants age $45-59$ (MiddleAge), 117 participants age $60-74$ (YoungOld), and 62 participants age 75+ (OldOld). Of these participants, 16 individuals were diagnosed with dementia, 51 with MCI, and the rest were classified as cognitively healthy. For the inclusion and exclusion criteria for participants, refer to Tables 2.1 and 2.2. Participants took 4 minutes on average to complete each activity while the testing session for eight activities lasted approximately 1 hour.

Before beginning each of the 8 IADL activities in the smart home, participants were familiarized with the apartment layout (e.g., kitchen, dining room, living room)

and the location of closets and cupboards. Materials needed to complete the activities were placed in their most natural location. For instance, in the sweeping task a broom was placed in the supply closet and the medication dispenser along with cooking tools were placed in the kitchen cabinet.

As participants completed the activities, two examiners remained upstairs in the apartment, watching the activities through live feed video. As the participant completed the activities, the examiners observed the participant and recorded the actions based on the sequence and accuracy of the steps completed. The experimenters also recorded extraneous participant actions (e.g., searching for items in wrong locations). Experimenter-based direct observation scores were later assigned by two coders who had access to the videos. The coders were blind to diagnostic classification of the older adults. Each activity was coded for six different types of errors: critical omissions, critical substitutions, non-critical omissions, non-critical substitutions, irrelevant actions and inefficient actions. The scoring criteria listed in Table 4.1 were then used to assign a score to each activity. A correct and complete activity received a lower score, while an incorrect, incomplete, or uninitiated activity received a higher score. The final direct observation score was obtained by summing the individual activity scores and ranged from 8 to 32. Agreement between coders for the overall activity score remained near 95% across each diagnostic group, suggesting good scoring reliability.

Figure 4.1 shows the distribution of the direct observation scores grouped by

Table 4.1: Coding scheme to assign direct observation scores to each activity.

Score	Criteria
1	Task completed without any errors
2	Task completed with no more than two of the following errors: non-critical omissions, non-critical substitutions, irrelevant actions, inefficient actions
3	Task completed with more than two of the following errors: non-critical omissions, non-critical substitutions, irrelevant actions, inefficient actions
4	Task incomplete, more than 50% of the task completed, contains critical omission or substitution error
5	Task incomplete, less than 50% of the task completed, contains critical omission or substitution error

participant age and cognitive classification. As participants completed the activities, the examiners recorded the time each subtask began and ended. These timings were later confirmed by watching video of the activity. Using this information, a research team member annotated raw sensor events in the data with the label of the subtask that the individual was performing when the event was triggered. Figure 4.2 shows a sample of the collected raw and annotated sensor data.

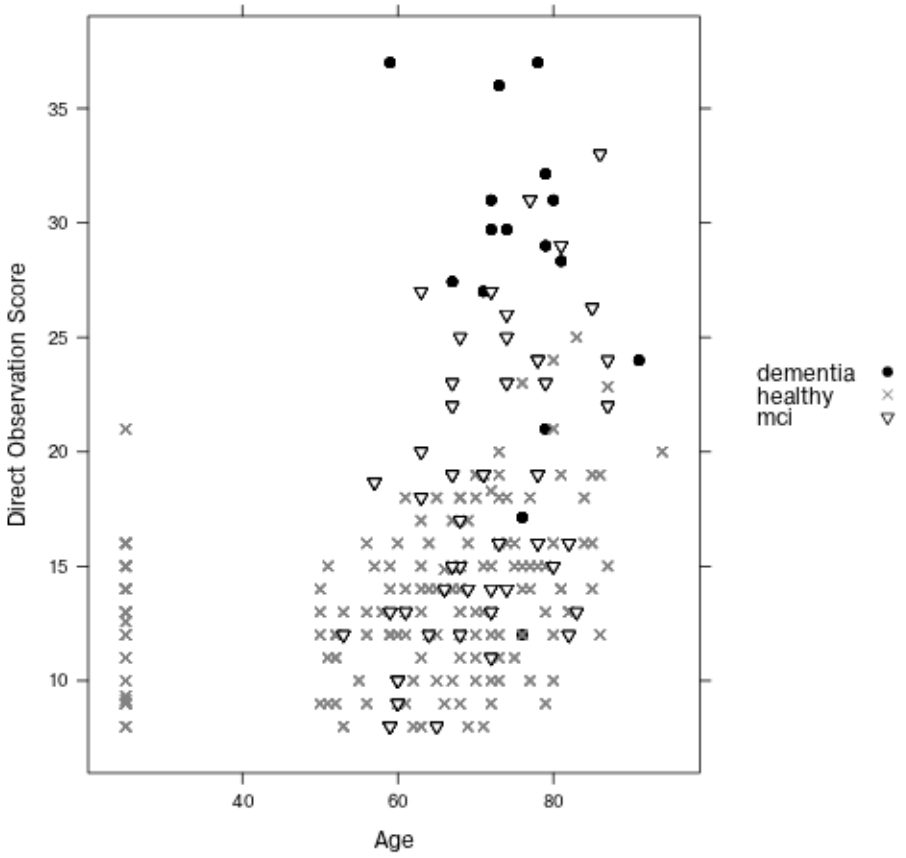


Figure 4.1: Distribution of the direct observation scores grouped by participant's cognitive diagnosis. Participants are organized by age on the x-axis and the y-axis represents the corresponding score.

4.2 Extracting features from smart home sensor data

We define features that can be automatically derived from the sensor data, reflect activity performance, and can be fed as input to machine learning algorithms to quantify activity quality and assess cognitive health status. These features capture salient information regarding a participant's ability to perform IADLs. Table 4.2 summarizes the 35 activity features that our computer program extracts and uses as input to the machine learning algorithms. The last feature, Health status, represents the target class label that our machine learning algorithm will identify based on the feature values.

As Table 4.2 indicates, we included the age of the participant as a discriminating feature because prior research showing age-related effects on the efficiency and quality of everyday task completion [131, 160] .

During task completion, participants sometimes requested help from the experimenter triggering a microphone, and this additional help is noted as a feature value. The experimenters also assigned poorer observed activity quality ratings when the participant took an unusually long time to complete the activity, the participant wandered while trying to remember the next step, the participant explored a space repeatedly (e.g., opened and shut a cabinet door multiple times) as they completed a step, or the participant performed a step in an incorrect manner (e.g., used the wrong

Date	Time	ID	Message
2010-01-27	09:04:07.001645	D007	OPEN Medicine-start
2010-01-27	09:04:07.004057	M016	OFF
2010-01-27	09:04:07.006439	M013	OFF
2010-01-27	09:04:09.009116	I006	ABSENT Medicine-step 2
2010-01-27	09:04:10.009602	M017	OFF
2010-01-27	09:04:11.058125	M017	ON
2010-01-27	09:04:13.004011	M017	OFF
2010-01-27	09:04:26.095202	I010	ABSENT Medicine-step 1
2010-01-27	09:04:29.006123	M017	ON Medicine-step 2
2010-01-27	09:04:31.050852	M017	OFF
2010-01-27	09:04:33.005186	P001	3439
2010-01-27	09:04:45.005352	P001	3428
2010-01-27	09:04:54.059162	M017	ON Medicine-step 2
2010-01-27	09:04:56.008738	M017	OFF
2010-01-27	09:06:18.086175	I010	PRESENT
2010-01-27	09:06:20.011737	I010	ABSENT
2010-01-27	09:06:20.014137	I006	PRESENT Medicine-step 3
2010-01-27	09:06:24.046323	M017	ON Medicine-step 3
2010-01-27	09:06:25.001502	P001	1185
2010-01-27	09:06:26.000764	M017	OFF
2010-01-27	09:06:29.091058	I010	PRESENT Medicine-step 3
2010-01-27	09:06:33.000376	D007	CLOSE Medicine-end

Figure 4.2: Sensor file format and sample annotation. Sensor IDs starting with M represent motion sensors, D represents door sensors, I represents item sensors, and P represents power usage sensors. The data is annotated with the start and end points of the activity (in this case, medicine) and the individual step numbers within the activity.

tool). The smart features are designed to capture these types of errors. The length of the event is measured in time (duration) and in the length of the sensor sequence that was generated (sequence length).

To monitor activity correctness, the number of unique sensor identifiers that triggered events (sensor count) is captured as well as the number of events triggered by each individual sensor (motion, door, and item sensor counts). Finally, for each activity the smart home software automatically determines the sensor identifiers that are related to the activity, or are most heavily used in the context of the activity, by determining the probability that they will be triggered during the activity. The sensors that have a probability greater than 90% of being triggered (based on sample data) are considered related to the activity the rest are considered unrelated. Therefore, the number of unrelated sensors that are triggered is noted as well as the number of sensor events caused by these unrelated sensors while a participant is performing the activity. These errors can widely be categorized into four different categories [38].

- Omission error: Omission error occurs when a participant does not perform an important subtask (e.g., failing to retrieve broom for sweeping task).

- Substitution error: Substitution error occurs when a participant uses an alternate object or incorrect gesture while completing the activity (e.g., dusting the living room instead of the kitchen).

- Irrelevant action: Irrelevant action happens when a participant carries out an action that is unrelated to the activity (e.g., opening a cupboard door while sweeping the kitchen).

- Inefficient action: Inefficient action happens when a participant carries out an inefficient task (e.g., opening extraneous cupboard doors).

Our feature extraction method does not consider individual "activity steps" while extracting features from the activities. As a result, the features are generalizable to any activity and not fine-tuned to the characteristics of a particular task. This means that the method does not have to be fine-tuned for a particular activity and its steps, but rather will consider features of any activity as a whole. As a result, the technique will be more generalizable to new activities. In addition, it is sometimes difficult to differentiate activity steps from environmental sensors. For example, it is difficult to detect individual steps of the outfit selection activity (moving to the closet, choosing and outfit, and laying out clothes) using only motion sensor data.

The list of features shown in Table 4.2 is extracted for all eight activities. Our machine learning algorithm receives as input a list of values for each of these 35 sensor-derived features and learns a mapping from the feature values to a target class value (health status). In order to train the algorithm and validate its performance on unseen data, ground truth values are provided for the participants in our study. Ground truth data for a participant is generated from a comprehensive clinical assessment,

which includes neuropsychological testing data (described previously), interview with a knowledgeable informant, completion of the clinical dementia rating [80, 97], the telephone interview of cognitive status [18], and a review of medical records. Figure 4.3 highlights the steps of the automated task assessment.

We observe that participants with cognitive disabilities often leave activities incomplete. Features of incomplete activities as thus denoted as missing. In the final dataset, we only include participants completing 5 or more activities (more than half of the total activities). The final dataset contains 47(2%) missing instances.

4.3 Automated task assessment

4.3.1 Method

The first goal is to use machine learning techniques to provide automated activity quality assessment. Specifically, machine learning techniques are employed to identify correlation between our automatically-derived feature set based on smart home sensor data and the direct observation scores. To learn a mapping from sensor features to activity scores, two different techniques are considered: a supervised learning algorithm using a support vector machine (SVM) [140] and an unsupervised learning algorithm using principal component analysis (PCA) [87]. Support vector machines are supervised learning algorithms that learn a concept from labeled train-

Table 4.2: Sensor-based feature descriptors for a single activity.

Feature #	Feature Name	Feature Description
1	Age	Age of the participant
2	Help	An indicator that experimenter help was given so that the participant could complete the task
3	Duration	Time taken (in seconds) to complete the activity
4	Sequence length	Total number of sensor events comprising the activity
5	Sensor count	The number of unique sensors (out of 36) that were used for this activity
6...31	Motion sensor count	A vector representing the number of times each motion sensor was triggered (there are 26 motion sensors)
32	Door sensor count	Number of door sensor events
33	Item sensor count	Number of item sensor events
34	Unrelated sensors	Number of unrelated sensors that were triggered
35	Unrelated sensor count	Number of unrelated sensor events
36	Health status	*Status of the patient*: Healthy, MCI, or Dementia

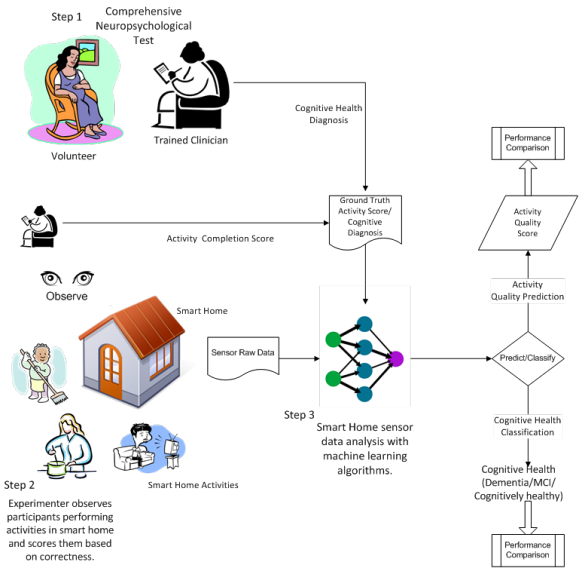


Figure 4.3: Steps involved in performing sensor-assisted cognitive health assessment. The process starts with a comprehensive neuropsychology assessment of the participant. The participant then performs IADLs in a smart home monitored by trained clinicians and smart home environmental sensors. The raw sensor data is annotated with activity labels. From the annotated sensor data, we extract features and analyze it with machine learning algorithms to derive the quality of the activity. The results are used by a clinician or by a computer program to perform cognitive health assessment.

ing data. They identify boundaries between classes that maximize the size of the gap between the boundary and the data points. A one vs. one support vector machine paradigm is used which is computationally efficient when learning multiple classes with possible imbalance in the amount of available training data for each class.

For an unsupervised approach, PCA is used to model activities. PCA is a linear dimensionality reduction technique that converts sets of features in a high-dimensional space to linearly uncorrelated variables, called principal components, in a lower dimension such that the first principal component has the largest possible variance, the second principal component has the second largest variance, and so forth. PCA is selected for its widespread effectiveness for a variety of domains. However, other dimensionality reduction techniques could also be employed for this task.

The eight activities used for this study varied dramatically in their ability to be sensed, in their difficulty, and in their likelihood to reflect errors consistent with cognitive impairment. Therefore, instead of learning a mapping between the entire dataset for an individual and a cumulative score, we build eight different models, each of which learns a mapping between a single activity and the corresponding direct observation score. Because the goal is to perform a direct comparison between these scores and the direct observation scores, and because the final direction observation scores represent a sum of the scores for the individual activities, the score output from our algorithm is also a sum of the eight individual activity scores generated by

the eight different learning models.

4.3.2 Experimental results

Empirical testing is to evaluate the automated activity scoring and compare scores with those provided by direct observation. The objective of the first experiment is to determine how well an automatically-generated score for a single activity correlates with the direct observation score for the same activity. In the second experiment, similar correlation analyses are performed to compare the automatically-generated combined score for all activities with the sum of the eight direct observation scores. In both cases support vector machines with bootstrap aggregation are used to output a score given the sensor features as input.

In addition to these experiments, a third experiment is performed to compare the automatically-generated combined score with the sum of the eight direct observation scores without using demographic and experimenter-provided features (age, was help provided) in the feature set. This provides greater insight on the role that only sensor information plays in automating task quality assessment. The bootstrap aggregation improves performance of an ensemble learning algorithm by training the base classifiers on randomly-sampled data from the training set. The learner averages individual numeric predictions to combine the base classifier predictions and generates

an output for each data point that corresponds to the highest-probability label.

Table 4.3 lists the correlation coefficient between automated scores and direct observation scores for individual activities and selected participant groups (cognitive healthy, MCI, and dementia) derived using SVM models. We note that correlation scores are stronger for activities that took more time, required a greater amount of physical movement and triggered more sensors such as Sweep as compared to activities such as Financial management. For activities like Financial management, errors in activity completion were more difficult for the sensors to capture. Thus, the correlation scores between automated sensor-based scores and direct observation scores in these activities are low. Similarly, we note that the correlation score also varies based on what groups (cognitively healthy, MCI, dementia) of participants are included in the training set. In almost all activities, the correlation is relatively strong when the training set contains activity sensor data for all three cognitive groups of participants.

Next, a combination of all of the performed activities is considered. Table 4.4 lists the correlation between a sum of the individual activity scores generated by the eight activity SVM models and the direct observation scores. Correlations between the two variables are relatively strong when the learning algorithm is trained using data from all three cognitive groups. Differences in correlation strength may be attributed to diversity present in the data. A majority of the cognitive healthy

Table 4.3: Pearson correlation between activity sensor-based scores and activity direct observation scores for sample subsets using SVM. For each sample subset, there are eight different learning models, each of which learns a mapping between a single activity and the corresponding direct observation score. The samples are cognitive healthy (CH) participants, participants with mild cognitive impairment (MCI) and participants with dementia (D) ($*p < 0.05$, $**p < 0.005, \dagger p < 0.05$ with Bonferroni correction for the three sample groups).

Pearson Correlation coefficient (r)									
Sample	N	Sweep	Medicine	Card	DVD	Plants	Phone	Cook	Dress
CH	196	0.50*†	0.02	0.04	0.22**†	0.04	0.31**†	0.18*†	0.22**†
MCI	51	0.58**†	0.01	0.07	0.13	0.01	0.18	0.35*†	0.03
CH,MCI	247	0.58**†	0.08	0.12	0.24**†	0.05	0.33**†	0.31**†	0.24**†
CH,D	212	0.58**†	0.16*†	0.09	0.24**†	0.08	0.31**†	0.28**†	0.22**†
MCI,D	67	0.75**†	0.01	0.21	0.03	0.38*	0.02	0.32**†	0.05
CH,MCI,D	263	0.63**†	0.17*†	0.07	0.27**†	0.09	0.33**†	0.37**†	0.23**†

participants completed the eight tasks correctly so the training data from this group contains examples of only "well-performed" activities and thus exhibits less diversity. Learning algorithms tend to generalize poorly when data contains little variation and thus classification performance may degrade.

Table 4.4 also lists the correlation between a sum of individual activity scores generated by the eight SVM models and the sum of the direct observation score without considering non-sensor-based features (age, help provided). Both Pearson linear correlation and Spearman rank correlation coefficient are calculated to assess relationship between variables. The correlation coefficients are statistically significant when correlations are derived only from the sensor-based feature set and duration does improve the strength of the sensor-based correlation. We conclude that demographic and experimenter-based features do contribute toward the correlation, but a correlation does exist as well between purely sensor-derived features and the direct observation score.

Tables 4.5 and 4.6 list the correlation coefficients between our automated scores and direct observation scores when we utilize PCA to generate the automated scores based on sensor features. Similar to the results in Tables 4.3 and 4.4, some activities have much stronger correlations than others and the strength of the correlations varies based on which groups are included in the training set. Furthermore, the correlation scores obtained using PCA are statistically significant but not as strong as

Table 4.4: Pearson correlation and Spearman rank correlation between the summed sensor-based scores and direct observation scores for sample subsets using SVM. Samples are cognitive healthy (CH) participants, participants with mild cognitive impairment (MCI), and participants with dementia (D) ($*p < 0.05$, $**p < 0.005, \dagger p < 0.05$ with Bonferroni correction for the three sample groups). The first correlation coefficient listed is the Pearson correlation coefficient while the second value is the Spearman rank correlation coefficient.

Sample	N	r (all features)	r (sensor features)	r (sensor features without duration)
CH	196	0.39**† 0.42**†	0.22**† 0.25**†	0.20** 0.22**
MCI	51	0.50**† 0.48**†	0.35* 0.31*	0.26* 0.20
D,CH	212	0.50**† 0.48**†	0.47**† 0.39**†	0.46**† 0.39**†
MCI,CH	247	0.49**† 0.48**†	0.34**† 0.32**†	0.32**† 0.30**†
MCI,D	67	0.59**† 0.63**†	0.60**† 0.60**†	0.53**† 0.52**†
CH,MCI,D	263	0.54**† 0.52**†	0.51**† 0.44**†	0.49**† 0.43**†

The column r (all features) lists correlation coefficients obtained using all features, r (sensor features) lists correlation coefficients obtained using only sensor-based features, and r (sensor features without duration) lists correlation coefficients obtained using all sensor-based features without the duration feature.

those obtained from the SVM models. Given the nature of the activities and given that the dimension of sensor derived features is reduced to a single dimension using a linear dimensionality reduction technique, it is likely that during the process some information is lost that otherwise produces a satisfactory correlation performance between direct observation scores and sensor-based features. Note that experiments are not performed which involve only participants from the dementia group because the sample size is small. Table 4.6 list the correlation coefficients between the automated scores and direct observation scores when PCA is utilized to generate the automated scores based on sensor features excluding non-sensor-based features. As before, we note that there is little difference between the sets of correlation coefficients.

These experiments indicate that it is possible to predict smart home task quality using smart home-based sensors and machine learning algorithms. We observe moderate correlations between direct observation score, which is a task quality score assigned by trained clinical coders, and an automated score generated from sensor features. We also note that the strength of the correlation depends on the diversity and quantity of training data. Finally, we also note that apart from the age of the participants, all of the features that are input to the machine learning algorithm are automatically generated from smart home sensor events.

Table 4.5: Correlation between activity sensor-based scores and activity direct observation scores for sample subsets using principal component analysis (PCA). We standardize the data before applying PCA. Samples are cognitive healthy (CH) participants, participants with mild cognitive impairment (MCI), and participants with Dementia (D) ($^*p < 0.05, ^{**}p < 0.005, \dagger p < 0.05$ with Bonferroni correction for the three sample groups).

Pearson Correlation coefficient (r)									
Sample	Size(N)	Sweep	Medicine	Card	DVD	Plants	Phone	Cook	Dress
CH	196	-0.30**†	0.09	0.06	0.23**†	0.04	0.03	0.26**†	-0.16*
MCI	51	-0.51**†	0.14	0.08	-0.17	0.38*†	0.20	0.11	-0.15
CH,D	212	-0.37**†	0.26**†	-0.02	0.22**†	0.06	0.04	0.18**†	-0.14*
CH,MCI	247	-0.36**†	0.13	0.00	0.25**†	0.20**†	0.10	0.23**†	0.08
MCI,D	67	-0.58**†	0.18	0.15	-0.11	0.20	0.19	-0.03	0.12
CH,MCI,D	263	-0.41**†	0.23**†	-0.05	0.23**†	0.15*†	0.10	0.17*†	0.07

Table 4.6: Pearson correlation and Spearman rank correlations between the summed sensor-based scores and direct observation scores for sample subsets using principal component analysis. We standardize data before applying PCA. The samples are cognitive healthy (CH) participants, participants with mild cognitive impairment (MCI), and participants with dementia (D) ($^*p < 0.05$, $^{**}p < 0.005, \dagger p < 0.05$ with Bonferroni correction for the three sample groups). The first value listed is the Pearson correlation coefficient while the second value is the Spearman rank correlation coefficient.

Sample	Sample size	r (all features)	r (sensor features)	r (sensor features without duration)
CH	196	0.13 0.12	0.12 0.10	0.11 0.08
MCI	51	0.10 -0.03	0.08 -0.08	0.09 -0.08
D,CH	212	0.13 0.12	0.11 0.09	0.10 0.08
MCI,CH	247	0.18**† 0.08	0.16*† 0.07	0.15* 0.05
MCI,D	67	0.12 0.09	0.10 0.08	0.10 0.07
CH,MCI,D	263	0.16* † 0.07	0.14* 0.05	0.13 0.03

The column r (all features) lists correlation coefficients obtained using all features, r (sensor features) lists correlation coefficients obtained using only sensor-based features, and r (sensor features without duration) lists correlation coefficients obtained using all sensor-based features without the duration feature.

4.4 Automated cognitive health assessment

4.4.1 Method

The second goal of this work is to perform automated cognitive health classification based on sensor data that is collected while an individual performs all eight activities in the smart home testbed. Here, a machine learning method is designed to map the sensor features to a single class label with three possible values: cognitively healthy (CH), mild cognitive impairment (MCI), or dementia (D).

When sensor data that was gathered for the population is visualized (shown in Figure 4.4), we see the heterogeneity of the data as well as specific differences in activity performance across the eight selected activities. As a result, we hypothesize that a single classifier would not be able to effectively learn a mapping from the entire data sequence for an individual to a single label for the individual. This is because individual activities vary in terms of difficulty, duration, and inherent variance. Machine learning researchers use ensemble methods, in which multiple machine learning models are combined to achieve better classification performance than a single model [50]. Here, eight base classifiers are initially created, one for each of the activities, using both a non-linear classifier (in this case, a support vector machine learning algorithm) and a linear classifier (in this case, a logistic regression classifier).

We observe that there is a class imbalance in our training set for cognitive health

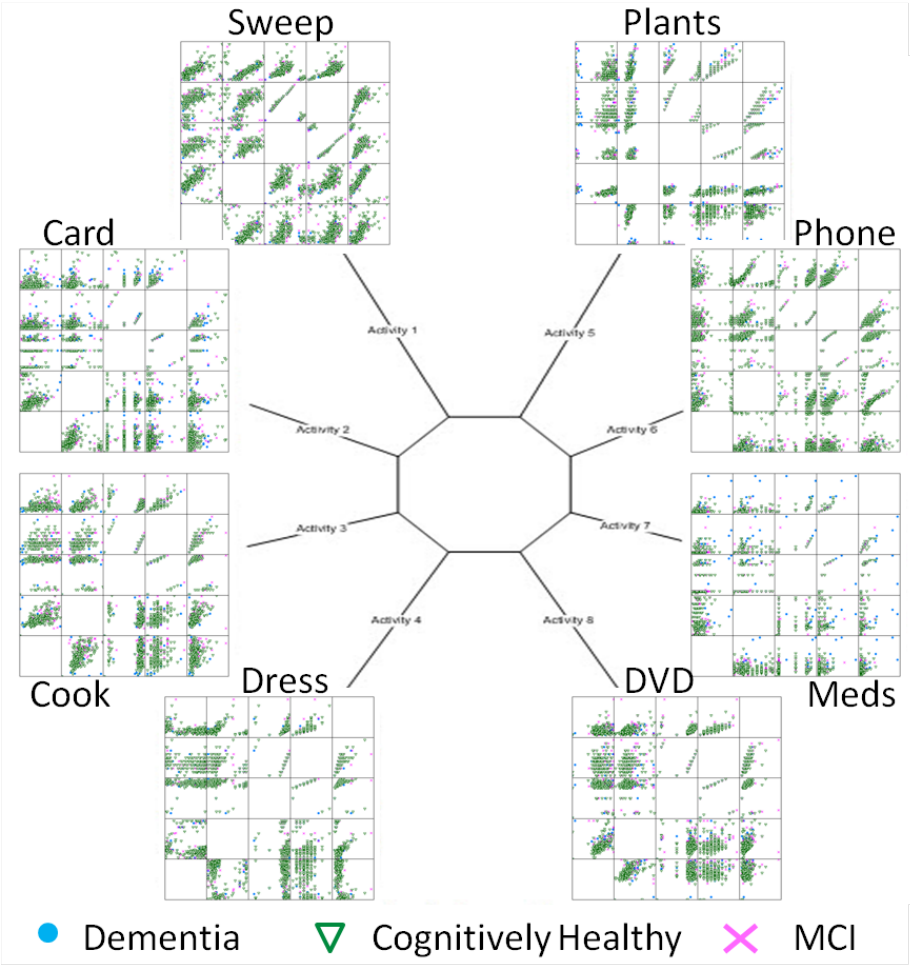


Figure 4.4: Scatter plot of sensor features for each of the eight activities. Each grid cell in the plot represents a combination of two of the sensor features (duration, sensor frequency, unrelated sensors, and unrelated sensor count).

prediction: there are only 16 individuals in the Dementia group and 51 in the MCI group, relative to the 196 participants in the cognitively healthy group. We note in

advance that such imbalance may adversely affect predictive performance as many classifiers tend to label the points with the majority class label. To address this issue, cost sensitive versions of the machine learning algorithms are used for each of the base classifiers. A cost sensitive classifier assigns misclassification costs separately for individual class labels and reweights the samples during training according to this cost. This allows the classifier to achieve overall strong performance even when the training points are not evenly divided among the alternative classes [145], as is the case with this dataset. A meta-classifier then outputs a label (CH, MCI, or D) based on a vote from the base learners.

4.4.2 Evaluation metrics

A number of evaluation metrics are utilized to validate the proposed methodology. The first, ROC curves, assess the predictive behavior of a learning algorithm independent of error cost and class distribution. The curve is obtained by plotting false positives vs. true positive at various threshold settings. The area under the ROC curve (AUC) provides a measure that evaluates the performance of the learning algorithm independent of error cost and class distribution.

In a data set with an imbalanced class distribution, g-mean measures the predictive performance of a learning algorithm for both the positive and a negative classes.

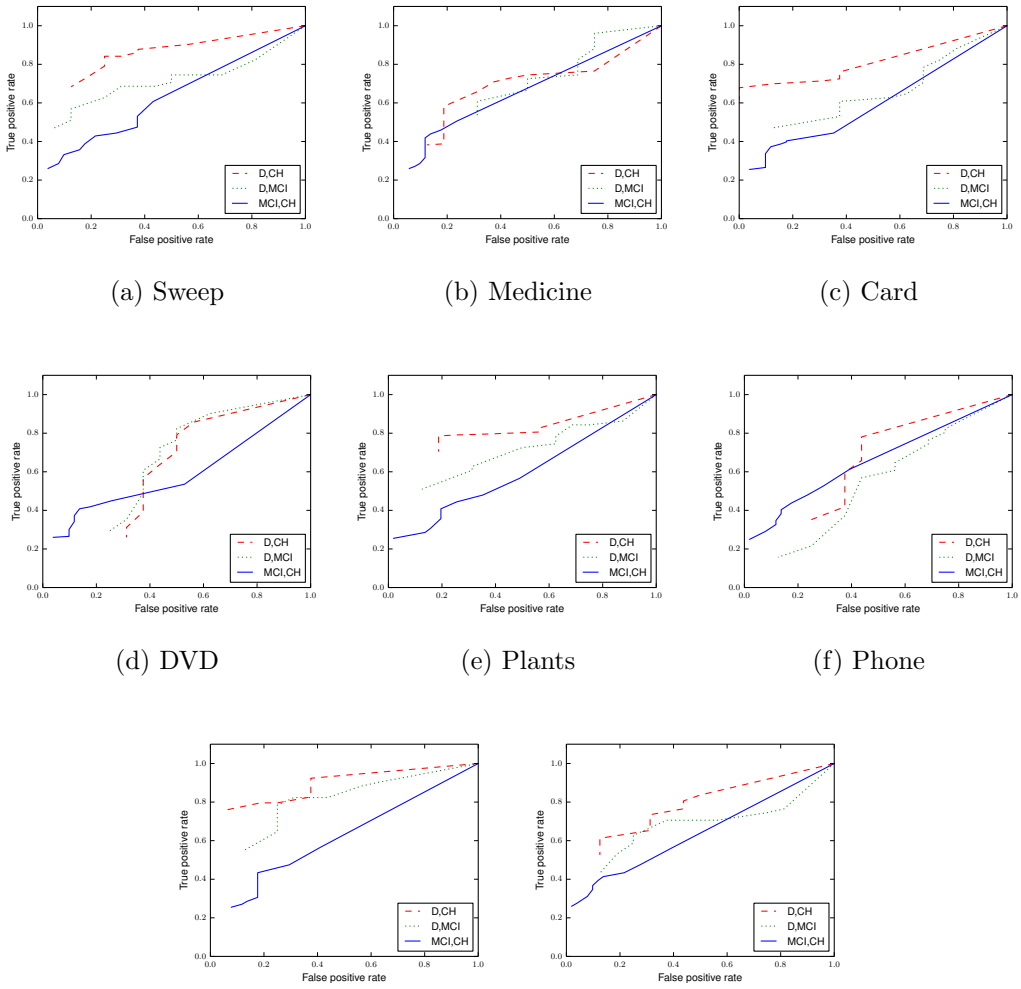
It is defined as:

$$gmean = \sqrt{(\text{true positive rate} \times \text{true negative rate})} \tag{4.1}$$

where the true positive rate and true negative rate represents the percentage of instances correctly classified to their respective classes. Furthermore, we also report if the prediction performance of a learning algorithm is better than random in both negative and positive classes. The classifier predicts a class better than random if the prediction performance, true positive rate, true negative rate, and the AUC value are all greater than 0.50.

4.5 Experimental results

Several experiments are performed to evaluate our automated cognitive health classifier. For all of the experiments we report performance based on overall area under the ROC curve (AUC) and g-mean scores. Values are generated using leave one out validation. To better understand the differences between each class, this situation is viewed as a set of binary classification problems in which each class is individually distinguished from another. The first experiment evaluates the ability of the classifier to perform automated health assessment using sensor information from individual activities using the support vector machine and logistic regression classifier algorithms.

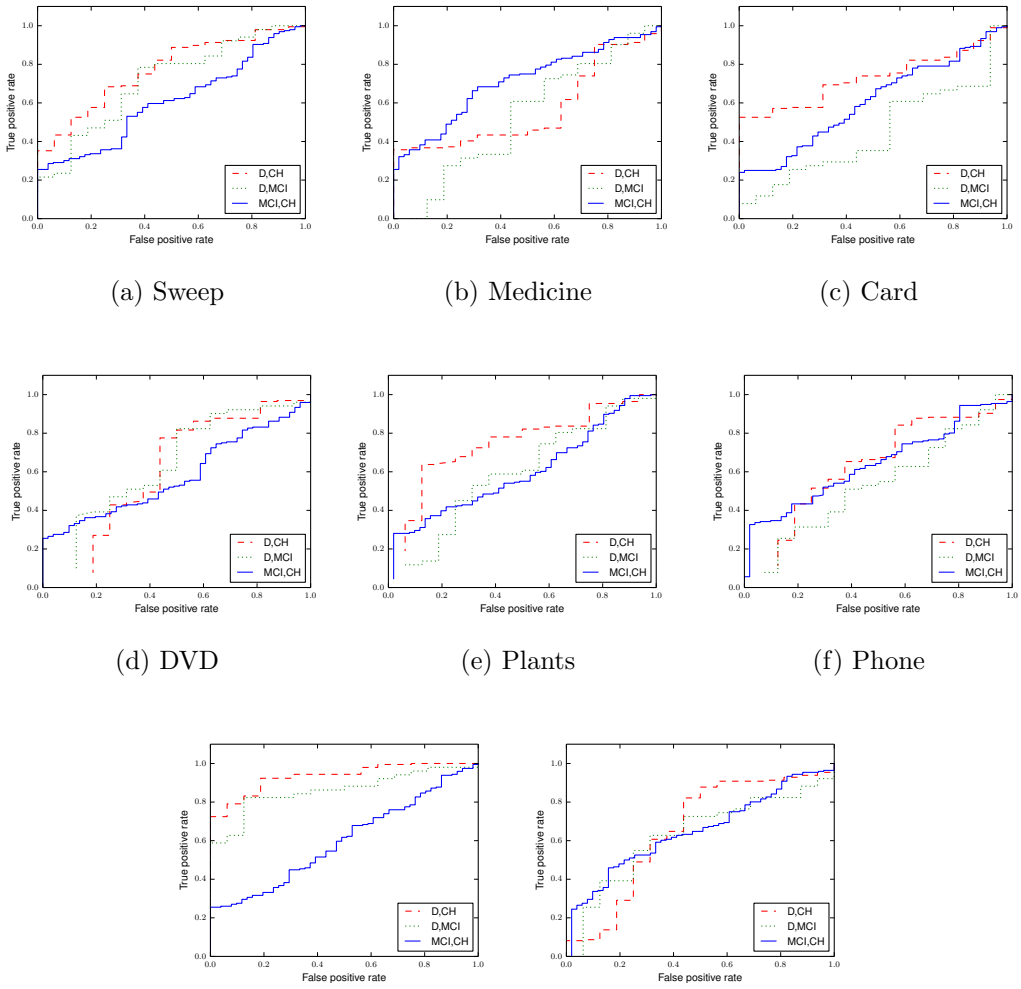


(a) Sweep (b) Medicine (c) Card

(d) DVD (e) Plants (f) Phone

(g) Cook (h) Dress

Figure 4.5: ROC curves for automated SVM classification of cognitive health status for each activity.



(a) Sweep (b) Medicine (c) Card

(d) DVD (e) Plants (f) Phone

(g) Cook (h) Dress

Figure 4.6: ROC curves for automated logistic regression classification of cognitive health status for each activity.

Table 4.7: AUC (first entry) and G-mean (second entry) values for automated support vector machine classification of cognitive health status for each activity.

Sample	Costs	Sweep	Medicine	Card	DVD	Plants	Phone	Cook	Dress
D, MCI	(3, 1)	0.70*	0.64	0.63	0.63*	0.68	0.54	0.78*	0.66*
		0.69	0.61	0.51	0.61	0.55	0.52	0.75	0.68
MCI,CH	(5, 1)	0.60	0.68	0.60	0.58	0.57	0.63	0.64	0.65
		0.57	0.64	0.62	0.57	0.57	0.62	0.56	0.59
CH,D	(23, 1)	0.82*	0.67*	0.81*	0.60*	0.76*	0.63	0.87*	0.76*
		0.79	0.69	0.81	0.60	0.79	0.57	0.73	0.69

*Classifier with better than weighted random prediction

Tables 4.7 and 4.8 and Figures 4.5 and 4.6 summarize the results. Classification performance for cognitively healthy vs. dementia is better than the other two cases. Similarly, performance for classifying MCI and cognitively healthy using SVM learning algorithm in all eight activities is not better than random prediction. Figure 4.1 shows that there is overlap between the direct observation scores of the healthy older adults and those diagnosed with MCI. MCI is often considered as a transition stage from cognitively healthy to dementia [97, 114]. It is possible that no reliably distinct differences exist between activity performances, or sensors are not able to capture

Table 4.8: AUC (first entry) and G-mean (second entry) values for automated logistic regression classification of cognitive health status for each activity.

Sample	Costs	Sweep	Medicine	Card	DVD	Plants	Phone	Cook	Dress
D, MCI	(3, 1)	0.71*	0.53	0.44	0.64	0.59	0.53	0.86*	0.63
		0.68	0.58	0.45	0. 62	0.58	0.54	0.75	0.61
MCI,CH	(5, 1)	0.61*	0.71*	0.60*	0.58	0.60*	0.64*	0.60	0.65*
		0.59	0.64	0.56	0.52	0.55	0.58	0.55	0.60
CH,D	(23, 1)	0.77	0.57	0.73	0.61	0.74	0.63	0.93*	0.66*
		0.66	0.48	0.58	0.63	0.50	0.57	0.80	0.67

*Classifier with better than weighted random prediction

subtle differences in activity performance between those two cognitive groups. Thus, additional experiments are not performed to distinguish between these two groups.

Similar to the results summarized in Section 5, we see that prediction performance for some activities such as Sweep, Dress, and Cook is better than for other activities such as DVD and Medicine. As explained previously, some of the activities took longer to complete and triggered more sensor events than others making it easier to identify errors, unrelated sensor events, and taking longer to perform the activity. Thus, differences exist in patterns of activity performance and sensors capture

Table 4.9: Combined cost-sensitive health classification performance with all activities classified using a SVM classifier.

Sample	Sample size	Costs	AUC	G mean
MCI,D	67	(2, 1)	0.56	0.43
MCI,CH	67	(2, 1)	0.62	0.59
D,CH	212	(5, 1)	0.72	0.65

Table 4.10: Combined cost-sensitive health classification performance with all activities classified using a logistic regression classifier.

Sample	Sample size	Costs	AUC	G mean
MCI,D	67	(2, 1)	0.53	0.40
MCI,CH	247	(5,1)	0.66	0.62
D,CH	212	(5, 1)	0.83	0.75

them. Our learning algorithm may be able to quantify these differences to distinguish between the different participant groups.

The second experiment evaluates the ability of the ensemble learner to automate health assessment using information from the combined activity set. The results are

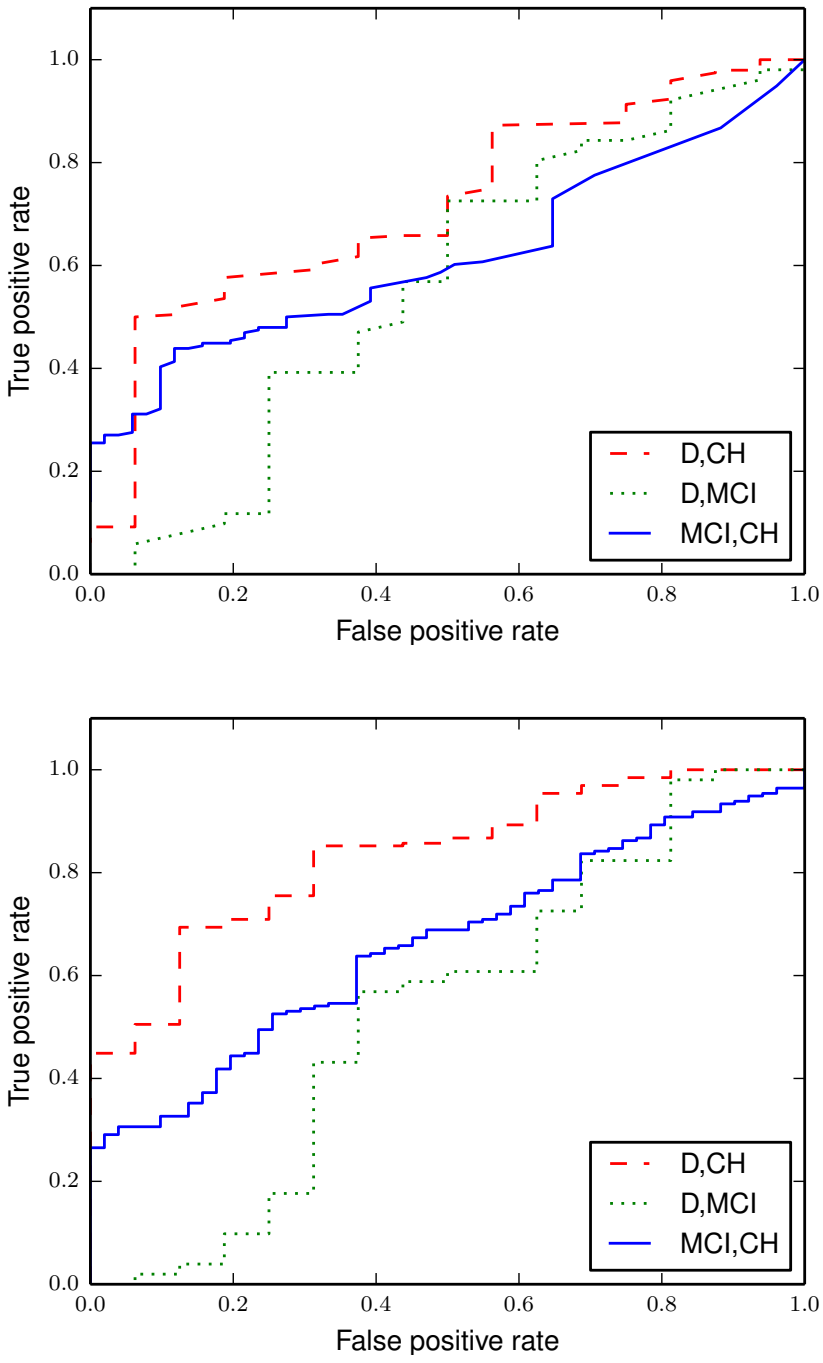


Figure 4.7: ROC curves for the combined cost-sensitive health classification with all activities classified using a SVM (top) and logistic regression classifier (down).

summarized in Tables 4.9 and 4.10 and Figure 4.7. The classification performance for classifying dementia and cognitively healthy is better than for classifying MCI vs. dementia. In addition, in Table 4.7 only a few base classifiers have better than random prediction. For each of these tables, costs are reported that yield the most promising results for the classifier. In a third experiment, only base classifiers that have better than random performance are selected. The results are summarized in Tables 4.11 and 4.12 and Figure 4.8. As shown in Table 4.11, for predicting MCI vs. dementia only 4 base classifiers are selected while for predicting dementia vs. cognitively healthy 5 base classifiers are selected. Similarly, for the results summarized in Table 4.12, only 2 base classifiers are selected for predicting dementia vs. cognitively healthy, while 6 base classifiers are selected for predicting MCI vs. dementia and 2 are selected for predicting dementia vs. cognitively healthy. The classification performance of MCI vs. dementia and dementia vs. cognitively healthy improves as compared to the previous two cases.

These experiments indicate that it is possible to perform limited automated health assessment of individuals based on task performance as detected by smart home sensors. The feature extraction technique described here along with the learning algorithm design achieves good performance at differentiating the dementia and cognitively healthy groups as compared to the other binary comparisons. This limitation might be due to the current coarse-grained resolution of the environment sensors

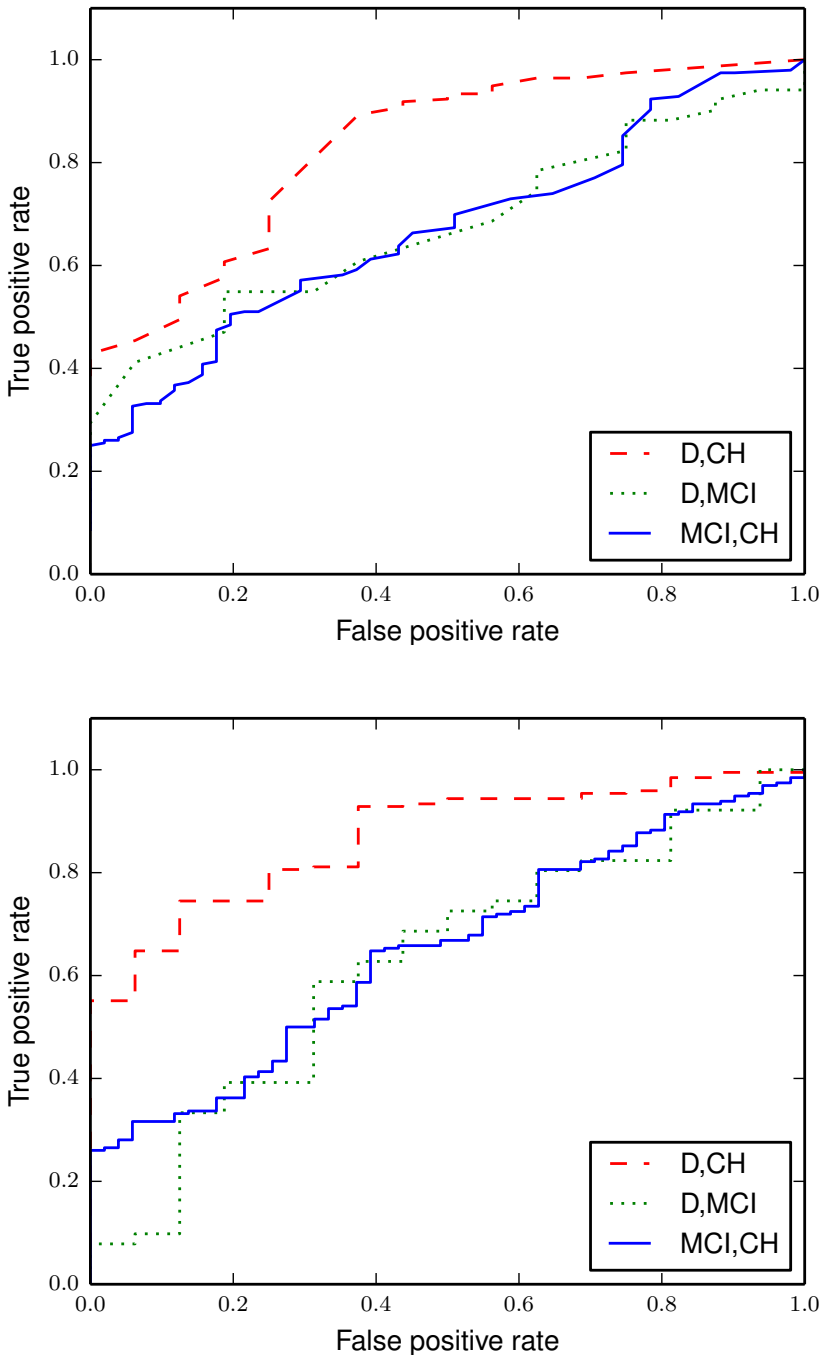


Figure 4.8: ROC curve for the combined cost-sensitive health classification with selected activities classified using a SVM (top) and logistic regression classifier (down).

Table 4.11: Combined cost-sensitive health classification with selected activities classified using a SVM classifier.

Sample	Sample size	Costs	AUC	G mean
MCI,D	67	(2, 1)	0.59	0.53
D,CH	212	(6, 1)	0.80	0.73

Table 4.12: Combined cost-sensitive health classification with selected activities classified using a logistic regression classifier.

Sample	Sample size	Costs	AUC	G mean
MCI,D	67	(3, 1)	0.62	0.54
MCI,CH	247	(3,1)	0.65	0.60
D,CH	212	(3, 1)	0.87	0.75

and the current smart home activity design. It might be possible to improve accuracy using different tasks, additional features, more sensors and those sensors that provide finer resolution such as wearable sensors.

4.6 Discussion and observations

In this chapter, we introduced a method to assist with automated cognitive assessment of an individual by analyzing the individual's performance on IADLs in a smart home utilizing a cross-sectional analysis. We hypothesize that learning algorithms can identify features that represent task-based difficulties such as errors, confusion, and wandering that an individual with cognitive impairment might commit while performing everyday activities. The experimental results suggest that sensor data collected in a smart home can be used to assess task quality and provide a score that correlates with direct observation scores provided by an experimenter. In addition, the results also suggest that machine learning techniques can be used to classify the cognitive status of an individual based on task performance as sensed in a smart home.

One must carefully interpret the correlation results that are mentioned here. The correlation (r) between smart home features and direct observation score is statistically significant. The correlation coefficient is squared to obtain the coefficient of determination. A coefficient of determination of 0.29 ($r = 0.54$) means that the nearly 30% of the variation in the dependent variable can be explained by the variation in the independent variable. The current results show that this method explains nearly 30% variations in the direct observational scores. Unexplained variation can

be attributed to limitations of sensor system infrastructures and algorithms.

We have seen that the predictive performance of a learning algorithm varies based on the activity being monitored and the condition of the individual performing the activities. As expected, the prediction accuracies of complex activities that triggered more sensor events were better than the accuracies for activities that trigger fewer sensor events and required less time to complete. Learning algorithms generalize better when trained from sensor rich data and when they are provided with data from a large segment of the population. The overall performance of a learning algorithm can also be improved by giving higher weights to the learning algorithms that model complicated activities and that take longer time to complete.

In addition, the prediction performance of the learning algorithm is affected by several factors. A primary factor is the class imbalance in our data set. Another contributing factor is missing values that are introduced in the cases when individuals (almost always individuals in the MCI and Dementia groups) do not attempt some of the activities. Finally, the ground truth values are based on human observation of a limited set of activities and may be prone to error. Based on these observations, we conclude that in a testing situation an experimenter needs to select activities with caution, balancing trade-off between a difficult activity that results in good prediction performance and one that is easy enough for participants with cognitive impairments to complete.

The current approach uses between-subjects differences in activity performance to perform cognitive health assessment and is based on set of non-obtrusive environmental sensors such as door sensors, motion sensors, and item sensors. The current work demonstrates that it is possible to automatically quantify the task quality of smart home activities and assist with assessment of the cognitive health of individual with a reasonable accuracy given the proper choice of smart home activities and appropriate training of learning algorithms.

In the next chapter, we will discuss another cross-sectional study in which participants completed a complex realistic activity in our smart home testbed. We present a machine learning algorithm to predict the activity quality of this complex activity utilizing sensor data that collected while participants perform this activity. This next study will allow us to look at activity performance across the population for a more complex, interweaved set of tasks than was discussed in this chapter.

CHAPTER 5. CROSS-SECTIONAL ANALYSIS OF DAY OUT TASK

In this chapter, we first discuss the design a smart home-based cross-sectional study in which participants performed a complex real lifelike activity, a Day Out Task (DOT), in our smart home testbed. We first introduce a machine learning-based method to assess the activity quality of DOT utilizing smart home sensor data collected while participants perform the activity. While participants perform eight different activities sequentially in the eight activities study, participants complete DOT activity by multitasking and parallelizing the subtasks of the DOT. Hence, DOT closely resembles the types of activities that humans perform in their daily life.

5.1 The testbed

Data is collected and analyzed using the Washington State University CASAS on-campus smart home testbed, an apartment that contains a living room, a dining area, and a kitchen on the first floor and two bedrooms, an office, and a bathroom on the second floor (see Section 2.1). A sample of the collected sensor events, together with the corresponding labeled activities, is shown in Figure 2.2.

Formally, the input data to our algorithm is a sequence of sensor events, E, that is generated as an individual performs an activity, A, which is comprised of subtasks $A_1 \ldots A_n$. A subtask A_i is represented by the corresponding sequence of n sensor events $e_1, \ldots, e_n$, the start time of the activity, the end time of the activity, and the activity label. Activity subtasks can be initiated in an arbitrary order and some activities or activity subtasks can be interwoven or parallelized. We state that activity A at sensor event e_i is parallelized if there is more than one subtask open (started but not ended) at that time.

5.2 The day out task

The ability to multi-task, or perform concurrent tasks or jobs by interleaving, has been said to be at the core of competency in everyday life [22]. We therefore designed a Day Out Task (DOT), a naturalistic task that participants complete by interweaving subtasks. Participants were told to imagine that they were planning for a day out, which would include meeting a friend at a museum at 10:00 AM and later traveling to the friend's house for dinner. The eight subtasks that need to be completed to prepare for the day out are explained and participants are told to multi-task and perform steps in any order to complete the preparation as efficiently as possible. Participants are also provided with a list and brief description of each

subtask that they can refer to during DOT completion. The eight subtasks are:

1. *Magazine*: Choose a magazine from the coffee table to read on the bus ride.

2. *Heating pad*: Microwave for 3 minutes a heating pad located in the kitchen cupboard to take on the bus.

3. *Medication*: Right before leaving, mime taking motion sickness medicine found in the kitchen cabinet.

4. *Bus map*: Plan a bus route using a provided map, determine the time that will be needed for the trip and calculate when to leave the house to make the bus.

5. *Change*: Gather correct change for the bus.

6. *Recipe*: Find a recipe for spaghetti sauce in a book and collect ingredients to make the sauce with a friend.

7. *Picnic basket*: Pack all of the items in a picnic basket located in the closet.

8. *Exit*: When all the preparations are made, take the picnic basket to the front door.

5.2.1 Experiment setup

Participants initially completed standardized and experimental neuropsychological tests in a lab. A Neuropsychology faculty member analyzed the test data to diagnose participants cognitive health. Participants in the dementia group met DSM-IV-TR criteria for dementia [10], which includes the presence of multiple cognitive deficits that negatively affect everyday functioning and represent a decline from a prior level of functioning. Inclusion criteria for MCI were consistent with the diagnostic criteria defined by Petersen [114, 115] and with criteria outlined by the National Institute on Aging-Alzheimer's Association workgroup [5]. We note that the clinical tests in DOT study and eight activities study are identical.

After completing the clinical tests, participants attempted the DOT task in our smart home testbed. While participants were completing the DOT, two experimenters (trained graduate students) remained upstairs in the apartment, watching participant performances through live feed video. As participant completed the DOT, the examiners recorded the time each subtask began and ended, events being interweaved, and subtasks goals being completed (e.g., retrieves magazine). As the individuals perform activities in the smart home, generated sensors events are recorded. Research team members (graduate students) annotated the sensor data to relate events with the label of the subtask that the individual was performing when the event was triggered.

Figure 2.2 shows a sample of the collected and annotated sensor data. Subtask accuracy scores and task sequencing scores were later assigned by coders after watching the video. Figure 4.3 illustrates this process.

To validate our approach for activity assessment, we include participants ($N = 179$) who completed at least two of the eight DOT subtasks. Among the participants included for analysis, 145 were cognitively healthy, 2 were diagnosed with dementia and 32 were diagnosed with MCI. We excluded 14 dementia participants who could not complete at least two DOT subtasks. The participant pool included 141 females and 38 males, with 37 ($N = 37$ CH) participants under 45 years of age (Young Young), 27 participants ($N = 4$ MCI, $N = 23$ CH) age 45-59 (MiddleAge), 84 ($N = 1$ dementia, $N = 20$ MCI, $N = 63$ cognitively healthy) participants age $60 - 74$ (YoungOld), and 31 ($N = 1$ dementia, $N = 8$ MCI, $N = 22$ CH) participants age 75+ (OldOld). The participants who completed only two subtasks took 10.4 ± 3.44 minutes in average to complete DOT while participants who completed all subtasks completed DOT in 9.83 ± 3.26 minutes to complete DOT. In average, participants took 10.33 ± 3.85 minutes to complete DOT task.

5.2.2 Task scoring

Two trained neuropsychologists watched the video data and, in conjunction with examiner-recorded data, assigned a task accuracy score and a sequencing score. The task accuracy score was based on the correctness and completeness of each of the eight subtasks. A correct and complete subtask received a lower score while an incorrect, incomplete, or uninitiated subtask received a higher score. The scoring criteria are listed in Tables 5.1 and 5.2. The final accuracy score was obtained by summing the individual scores of each task and thus ranged from 8 to 32. The task sequencing score represents whether the participant sequenced six of the DOT subtasks correctly. Participants received 1 point for each correct sequence (e.g., put the heating pad in the microwave for 3 minutes as one of the first four subtasks). The normalized range of scores is 1 to 6 such that lower score indicates a more correct and/or efficient sequencing of subtasks. Two coders, blinded to group assignment, independently assigned scores to participants based on specific criteria as they directly viewed the participants' task performance. Inter-rater reliability agreement for the accuracy and sequencing scores was 97.88% and 99.57%, respectively, and was calculated by dividing the number of responses by the total discrepancies due to double scoring [130]. Figure 5.1 shows the distribution of the direct observation scores, accuracy, and sequencing score grouped by participant cognitive diagnosis.

Table 5.1: Coding scheme to assign accuracy score to each subtask.

Accuracy score	Criteria
1	Complete / Efficient
2	Complete / Inefficient
3	Incomplete / Inaccurate
4	Never Attempted

Table 5.2: Coding scheme to assign sequencing score to each subtask. Total sequencing score is the count of yes responses to these criteria.

ID	Criteria
1	Heating pad started as one of first four activities.
2	Picnic basket retrieved as one of first four activities.
3	Cost of bus fare determined prior to first attempt at retrieving change.
4	Recipe read prior to retrieving first food item.
5	Motion Sickness pill taken near end.
6	Picnic basket moved to front door as one of last two activities.

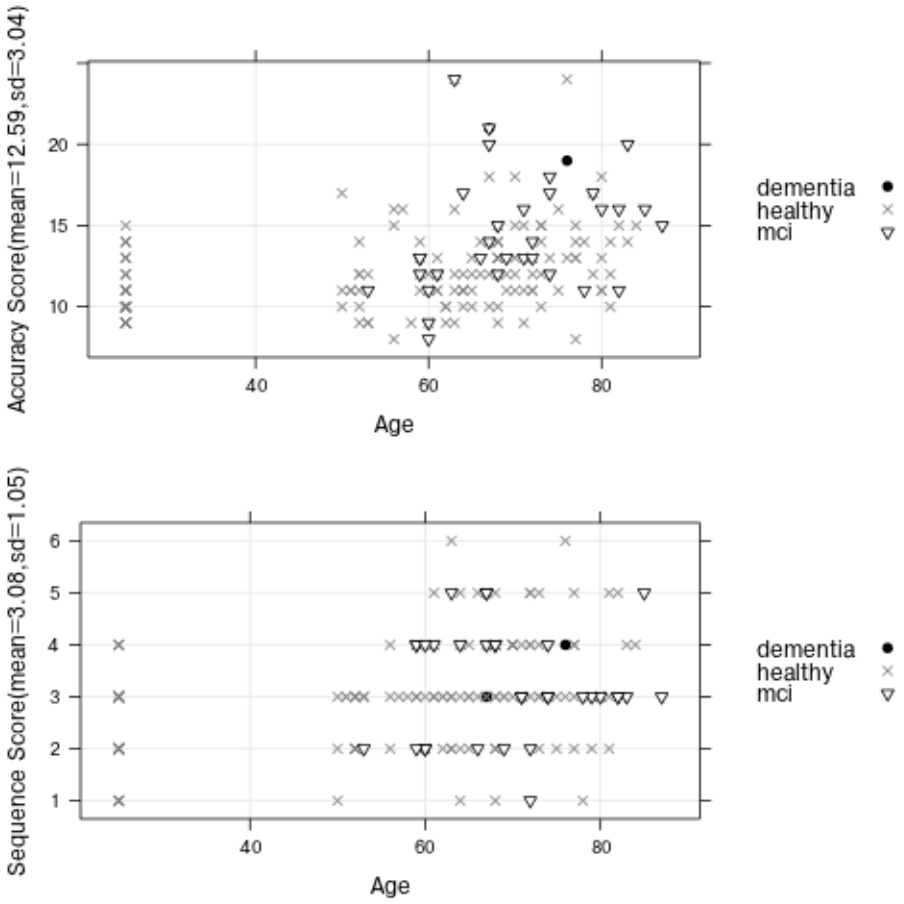


Figure 5.1: Distribution of the neuropsychologist direct observation scores, accuracy scores and sequencing scores, with participant's cognitive diagnosis indicated by point type. Individual participants are organized by age on the x-axis and by the corresponding score on the y-axis.

5.3 Feature extraction

To assess an individual's performance on the DOT, we derive features from sensor data that reflect task performance and can be input to a machine learning algorithm to quantify task quality. We define DOT performance based on the nature of activity completion and execution of the activity subtasks. A participant efficiently executes DOT if he multitasks DOT subtasks and sequences them correctly. Similarly, time taken to complete the entire DOT activity and number of sensors triggered during activity completion explains the participant's DOT performance. Features were chosen based on prior studies which found that, in comparison to cognitively healthy older adults, individuals with MCI complete everyday activities (e.g., locating nutrition information on food labels, making toast, medication management) more slowly and commit more errors, including errors of commission, omission [130], and task sequencing/tracking [136].We note that in this study the activity start points and end points were generated by human annotators. However, we can use activity recognition algorithms to automate this step [138]. These features are explained below.

5.3.1 Duration

We use the duration feature to represent the total wall clock time that the participant takes to complete the entire set of DOT activities. The time to complete an activity can indicate participant's age, mobility and overall cognitive health. If subtasks are executed independently then we can consider the time for each subtask as a separate feature. For the DOT, subtasks are interleaved and performed in parallel, so we consider time taken for the entire DOT.

5.3.2 Number of sensors and sensor events

This feature reflects the spatial areas and objects that are manipulated while DOT is being performed. The number of sensors indicates the number of different sensor identifiers that generate events during the DOT, while the number of sensor events keeps track of the number of events that is generated by each unique sensor in the space. These counts provide insight on the type of activities that are being performed and how well the participant stays on the task. For example, some participants wandered out of the normal activity region, used incorrect tools for a subtask, or explored the same space, cabinet or region repetitively as they attempted to complete the appropriate subtask.

5.3.3 Number of complete activities

Not all participants completed all DOT subtasks. We thus introduce an ActivitiesCompleted feature which indicates whether the participant completed all of the DOT subtasks.

5.3.4 Pattern sequencing

In the case of a complex activity such as the DOT, subtasks can be performed with many order variations. For instance, one participant might choose a magazine first, while another might start by first looking up a recipe. Participants are expected to parallelize subtasks for efficiency. However, some subtask sequences and parallelisms are more efficient than others. As an example, if a participant starts the DOT by microwaving a heating pad, they are able to complete other tasks while waiting for the microwave to finish. If they wait until the end of the DOT to microwave the heating pad this parallelism is not possible. We hypothesize that the sequence in which tasks are performed influences the amount of parallelism that can be achieved and thereby affects the efficiency of the overall task. To represent task sequencing choices, we define a DOT sequencing vector $s_1, s_2, .., s_8$ that encodes the order in which an individual started various tasks (in the DOT, there are 8 such tasks to choose from). For example, the sequencing vector $(2, 3, 1, 4, 5, 6, 7, 8)$ indicates that the 2^{nd} task in

the set was initiated first, followed by the 3^{rd} task, then the 1st task, and so forth. If an individual does not initiate a particular task, then the corresponding position in the vector sequence is treated as missing based on the sequences that were performed by others in the population.

5.3.5 Activity interruptions

In the case of activities that involve waiting for an event (e.g., waiting for the Heating Pad to warm up), interrupting the activity to finish other tasks is both efficient and is an indication that the participant is capable of generating more complex plans that interweave multiple activities. However, for activities that take a short time to complete such as Change and Bus/Map, participants will likely complete the task without interruptions. To capture differences in interruptions on various activities, we define activity interruption features based on all DOT subtasks. For long activities, such interruptions may indicate that the participant is able to generate a complex and efficient DOT solution.

5.3.6 Parallelism

Participants in our study were encouraged to multitask the DOT subtasks as much as possible to complete the DOT quickly. The ability to multitask varied

dramatically among individuals and was expected to present a challenge for those with dementia and MCI. We were therefore interested in quantifying the amount of parallelism or multitasking that existed in an individual's performance of the DOT.

To quantify parallelism, we introduce a variable called activity level, a_i, that represents the number of activities that are open (i.e., that have been started but not completed), at the time that sensor event e_i is generated. A set of activity levels $\{a_1, a_2, .., a_n\}$ can be defined for all of the sensor events that were generated during the DOT. To represent this set more succinctly we employ run length encoding (RLE). A run for an activity level is a string of equal-valued activity levels. RLE encodes runs of activity levels as activity levels with corresponding counts, as shown in Figure 5.2. Based on run length encoding, we derive a $M \times N$ run length matrix P, where M is the maximum activity level and N is the length of the sensor sequence. Each element of the matrix, $P(x, y)$, represents the number of runs of length y corresponding to activity level x, or the number of times that activity level x occurs y consecutive times. A similar technique has been used to analyze computed tomography volumetric data to capture various text characteristics [47].

We introduce two measures, the High Activity-Level Run Measure (HALRM) and the Low Activity-Level Run Measure (LALRM), to capture a participant's level of task parallelism that occurred over a sequence of sensor events. If a participant parallelizes subtasks for a longer period of time we expect his HALRM to be high,

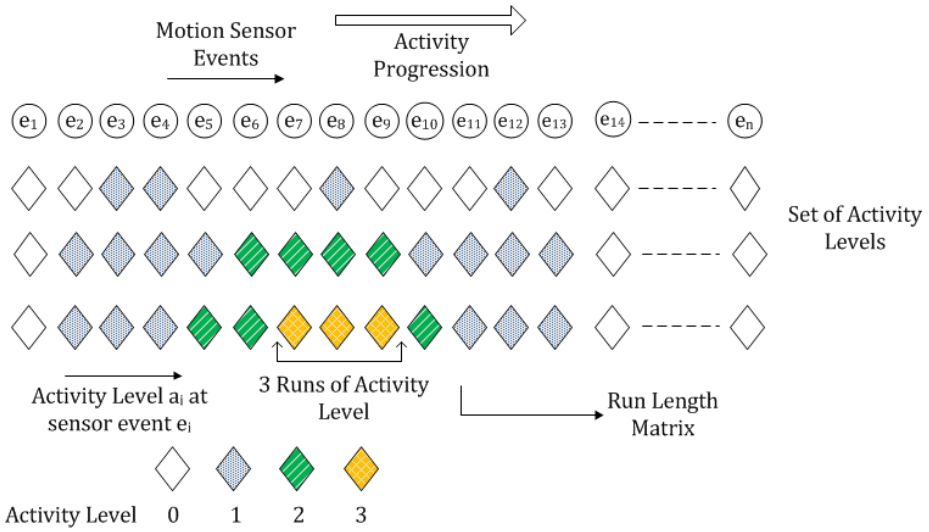


Figure 5.2: Sets of activity levels for three participants. The first item of the set represents activity level at the initial sensor event. As activity progresses, sets are augmented with activity levels for different sensor events. For example, during the eighth sensor event e8, participant III has activity level 3 and participant I has 1. The run length matrix takes this activity level set as input.

while if he does not parallelize subtasks his LALRM measure would be high.

$$HALRM = \sum_{i=1}^{M} \sum_{j=1}^{N} P(i,j) \times i \times j \tag{5.1}$$

$$LALRM = \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{P(i,j) \times j}{i} \tag{5.2}$$

Based on these two measures, we define the parallelizing index, Pindex, to represent the amount of task interweaving that is performed. Pindex is computed as the ratio of HALRM to LALRM, as shown in Equation 5.3.

$$Pindex = \frac{HALRM}{LALRM} \tag{5.3}$$

As Equation 5.3 indicates, a higher parallelizing index indicates a higher level of parallelism in the activity. It does not reflect a higher quality of DOT. For example, a participant may have a high Pindex because he initiated many of the subtasks. On the other hand, he may leave subtasks incomplete or take a long time to complete the subtasks. The Pindex does provide particularly useful insights on task quality when combined with the other task features.

We note that the both parallelism and interruption-based features are calculated using the DOT subtasks that were manually annotated in the sensor data. For example, in the parallelism feature, we calculate an activity level a_i which is the number of activities that are open (started but not yet completed) at the time sensor event

Table 5.3: DOT feature set.

Feature Set	Feature Type
DOT features	Duration, sensor counts, sensor events, activity completeness
Interruption features	Number of activity interruptions
Sequencing features	Sequence vector
Parallelism feature	Pindex

e_i is generated. To calculate a_i, we use all of the DOT subtasks that were annotated as "started" or "ended" in all of the previous sensor events from a sensor event e_i. Similarly, to calculate the interruption-based feature for a subtask (for example, the medication subtask), we use all of the sensor events and the activity annotations between "medication-start" and "medication-end" sensor events. We note that we can also use an activity recognition algorithm to automatically annotate these activity labels in the sensor data.

The set of extracted features is summarized and categorized in Table 5.3. We hypothesize that these smart home features will allow us to provide automated task quality scores that correlate with task scores obtained by direct observation.

5.4 Automated scoring

DOT task accuracy and task sequencing scores are derived from direct observation of participant's task performance. We used machine learning techniques to identify correlation between our automated feature set based on smart home sensor data and the direct observation scores. We describe two approaches to automated scoring: supervised learning and using unsupervised learning techniques.

5.4.1 Supervised scoring models

We formulate the automated scoring problem as a supervised learning problem in which a learning algorithm learns a function that maps the sensor-derived features to the direct observation scores. We use a support vector machine (SVM) with sequential machine optimization and bootstrap aggregation or bagging to learn the mapping. Support vector machines identify class boundaries that maximize the size of the gap between the boundary and data points. The bootstrap aggregation improves performance of an ensemble learning algorithm by training the base classifiers on randomly-sampled data from the training set. The learner averages individual numeric predictions to combine the base classifier predictions and generates an output for each data point that corresponds to the highest-probability label. We use both supervised regression and classification algorithms in our supervised scoring models.

5.4.2 Unsupervised scoring models

A score that is generated by a supervised learning algorithm predicts the quality of an activity in a way that emulates human-assigned scores. In contrast, unsupervised techniques use characteristics of the data itself to identify natural boundaries between activity performance classes. Here we derive unsupervised scores using a dimensionality reduction technique. Dimensionality reduction techniques reduce a high-dimensional dataset to one with a lower dimension. We use this to reduce the feature set to a single numeric score. While we use Principal Component Analysis (PCA) to reduce the dimension, many reduction techniques would be appropriate for this task [87]. PCA is a linear dimensionality reduction technique that converts sets of features in a high-dimensional space to linearly uncorrelated variables, called principal components, in a lower dimension such that the first principal component has the largest possible variance, the second principal component has the second largest variance, and so forth. After reducing the dimension, we use min-max normalization to convert the variables to a uniform range.

5.4.3 Cognitive assessment models

In our final step, we evaluate the use of smart home techniques to automate the cognitive health assessment of participants based on sensor-based features that

describe their activity performance. We map each participant to one of the three cognitive groups: Dementia (D), Mild Cognitive Impairment (MCI), or Cognitively Healthy (CH). To accomplish this, we extract the same sensor-based activity features that were used for the earlier experiments, as explained in Section 4. We obtain ground truth cognitive health labels for each participant from a battery of standardized and experimental neuropsychological tests that were administered in a clinical setting. We then train learning algorithms to learn a mapping from the sensor-based activity features to the cognitive health label (CH, MCI or D).

5.5 Evaluation

Our goal is to design smart home technologies to automate assessment of task quality and of cognitive health. We evaluate our approaches using data collected on a smart home testbed. We evaluate the two tasks separately. To evaluate the ability to automate assessment of task quality, we compare scores generated from our smart home algorithm with direct observation scores generated from neuropsychologists and to evaluate the ability to automate assessment of cognitive health, we compare diagnoses generated from our algorithms with diagnoses based on clinical tests.

We perform four experiments to evaluate our smart home-based task quality assessment algorithms. First, we measure the correlation between subsets of our smart

home sensor features and direct observation scores (Section 5.5.1) . Second, we measure the correlation between the entire set of sensor features and direct observation scores (Section 5.5.1). Third, we assess how well a SVM correctly classifies task quality, using the direct observation scores as ground truth labels (Section 5.5.1). Finally, we determine how well the scores derived using unsupervised algorithm correlates with direct observation scores (Section 5.5.2).

In addition, we evaluate learning algorithms using different participant groups that we construct based on their cognitive diagnosis (D, MCI, and CH) and number of subtasks they complete. Since the number of cognitively healthy participants is large, we further divide them to Older adults (Middle Age, Young Old, and Old Old) and Younger adults (Young Young). These sample groups have different heterogeneity. We refer to a sample group as heterogeneous if it contains examples of both well-conducted and poorly conducted activities.

Training set containing instances of cognitively healthy individuals who commit fewer mistakes tend to be less heterogeneous as compared to training set containing instances of both cognitively healthy individuals and individuals with MCI who often commit more mistakes. Similarly, individuals who complete fewer subtasks normally commit more mistakes than individuals who complete a higher number of subtasks. By training learning algorithm using these sample subsets, we can understand how the heterogeneity impacts the performance of the learning algorithms and helps us to

understand the features of these different groups.

We next evaluate the ability of our learning algorithm to map smart home activity sensor features to a cognitive health diagnosis. We train learning algorithms using smart home data and the cognitive health assessments provided by trained clinicians (Section C) and evaluate them using two metrics: the Area under the ROC curve (AUC) and the F-score.

ROC curves assess the predictive behavior of a learning algorithm independent of error cost and class distribution. We plot false positives vs. true positive at various threshold settings to obtain a ROC curve. The area under the ROC curve (AUC) provides a measure that evaluates the performance of the learning algorithm independent of error cost and class distribution [164]. Similarly, the F-score is the harmonic mean of the precision and recall and is defined as [164]:

$$\text{F-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{5.4}$$

5.5.1 Evaluation of supervised scoring models

Feature subset correlation

For our experiment, we consider alternative feature subsets summarized in Tables 4 and 5. For each subset, we generate the correlation coefficient between the feature values derived from smart home sensor data and the experimenter direct

observation scores (the accuracy score and sequencing score). In addition, we also analyze varying subsets of participants. Specifically, we consider subgroups of participants corresponding to the individuals with dementia (sample D), individuals with MCI (sample M), older adults (sample OA), and younger adults (sample YA). The objective of the experiment is to identify the correlation that exists between smart home task feature subsets for each participant and the activity quality score for the participant provided by trained clinicians and based on direct observation of the activity.

Table 5.4: Correlations between feature subsets, participant groups, and the accuracy direct observation score. Samples are D=Dementia, M=MCI, OA= Cognitively Healthy Older Adult, YA= Cognitively Healthy Younger Adult. ($^*p < .05$, $^{**}p < .005$, $\dagger p < 0.05$ with Bonferroni correction for n sample groups)

Correlation coefficient (r)							
Participant sample	{D,M,OA,YA}	{D,M,YA}	{OA,YA}	{YA}	{M,YA}	{M }	{M,YA}
Sample Size	179	177	145	37	69	32	140
DOT features	0.58**†	0.57**†	0.57**†	0.52**†	0.54**†	0.44*	0.55**†
Interruption features	0.31**†	0.32*†	0.25*	0.21	0.27	0.4	0.36*
Sequencing features	0.76**†	0.72**†	0.64**†	0.36*	0.78**†	0.79**†	0.68**†
Parallelism feature	0.39**†	0.39**†	0.18*	0.11	0.59**†	0.58**†	0.39**†

From Tables 5.4 and 5.5, we see that correlations between most of the feature

Table 5.5: Correlations between feature subsets, participant groups, and the sequencing direct observation score. Samples are D=Dementia, M=MCI, OA= Cognitively Healthy Older Adult, YA= Cognitively Healthy Younger Adult. (*$p < .05$, **$p < .005$, †$p < 0.05$ with Bonferroni correction for n sample groups.)

Correlation coefficient (r)							
Participant sample	{D,M,OA,YA}	{M,OA,YA}	{OA,YA}	{YA}	{M,YA}	{M}	{M,OA}
Sample Size	179	177	145	37	69	32	140
DOT features	0.10	0.01	0.21	0.21	-0.01	-0.27	-0.08
Interruption features	0.43**†	0.42**†	0.45**†	0.47**†	0.28*	0.22	0.34**
Sequencing features	0.46**†	0.42**†	0.50**†	0.20	0.30*	-0.12	0.38**†
Parallelism feature	0.12	0.13	0.03	0.32	0.29*	0.01	0.02

subsets and direct observation accuracy/sequencing scores are statistically significant. We find that the correlation between the smart home features and the observation-based accuracy score is stronger than the correlation with observation-based sequencing scores. A possible reason is that the task accuracy score quantifies the correctness and completeness of the eight DOT subtasks, which reflects the same type of information captured by smart home features. In contrast, the sequencing score quantifies how the DOT subtasks were sequenced, which is not as extensively captured by smart home features.

We find that feature subsets correlate differently with different training sample subsets. For instance, in Table 5.4 DOT features have stronger correlation with task accuracy score but parallelism feature has weak correlation when we train learning algorithms with cognitively healthy younger adult group (column YA). This indicates that a learning algorithm can better predict task accuracy with DOT features than parallelism features when training set contains examples of cognitively healthy individuals. Similarly, in Table 5.4 we see that the parallelism features correlates higher when sample subsets of training data contain individuals with MCI and younger adults (column M,YA) but does not when it contains cognitively healthy individuals (column YA) indicating that parallelism features can better represent differences between younger adults and MCI. Thus, we see that predictive power of a feature set depends on participant groups.

In addition, we visualize the relationship between selected feature types and the direct observation scores. Figure 5.3 plots the order in which subtasks were initiated within the DOT. As the figure shows, most participants placed Bus Map first in their sequence and almost all participants initiated the Exit subtask last. There is a fairly consistent choice of ordering among the subtasks for all participants, with the greatest variation occurring in positions 3, 6, and 7 of the sequence. We thus conclude that task sequencing plays an important role in such a complex activity as the DOT and should be analyzed as a part of overall task quality.

In a separate step, we plot the relationship between Pindex (the parallelism feature) and the direct observation scores. As shown in the left plot in Figure 5.4, Pindex consistently increases with accuracy score. The figure also shows a relationship between Pindex and the sequence score, although it is not as distinct. We note that when a participant initiates but does not complete subtasks their task quality degrades which increases their Pindex score. Correspondingly, as mentioned in Table 5.1, their accuracy score increases as well.

Combined feature correlation

In this experiment, we use the SVM regression and bootstrap aggregation to learn a regression model that finds a fit between the combined set of feature values and the accuracy and the sequencing direct observation score. There are two objectives of this experiment. The first objective is to evaluate the correlation between the smart

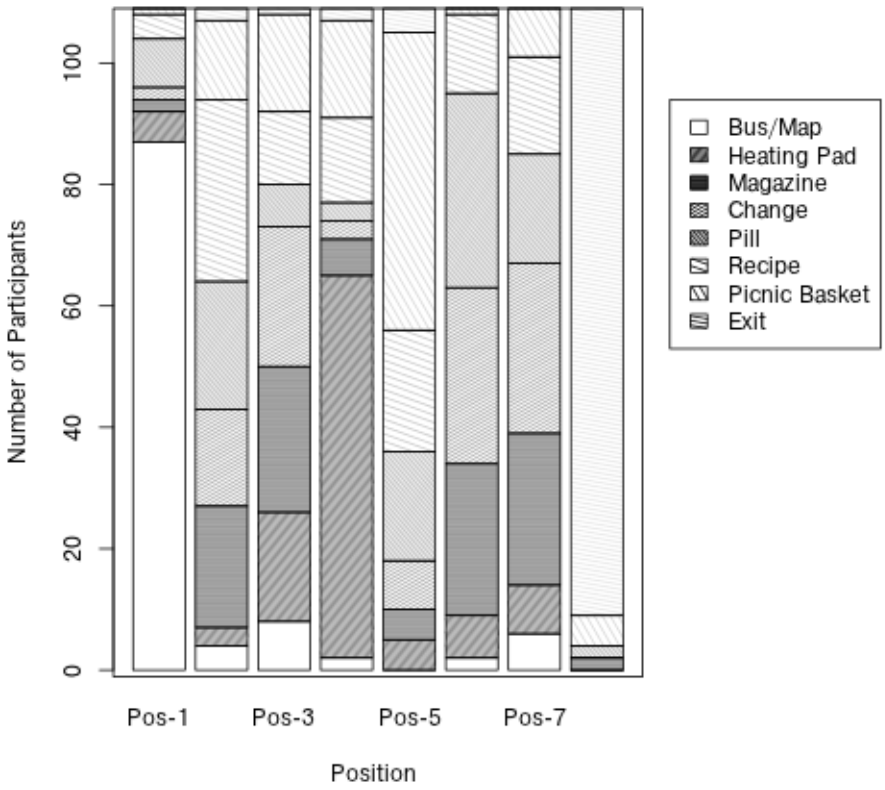
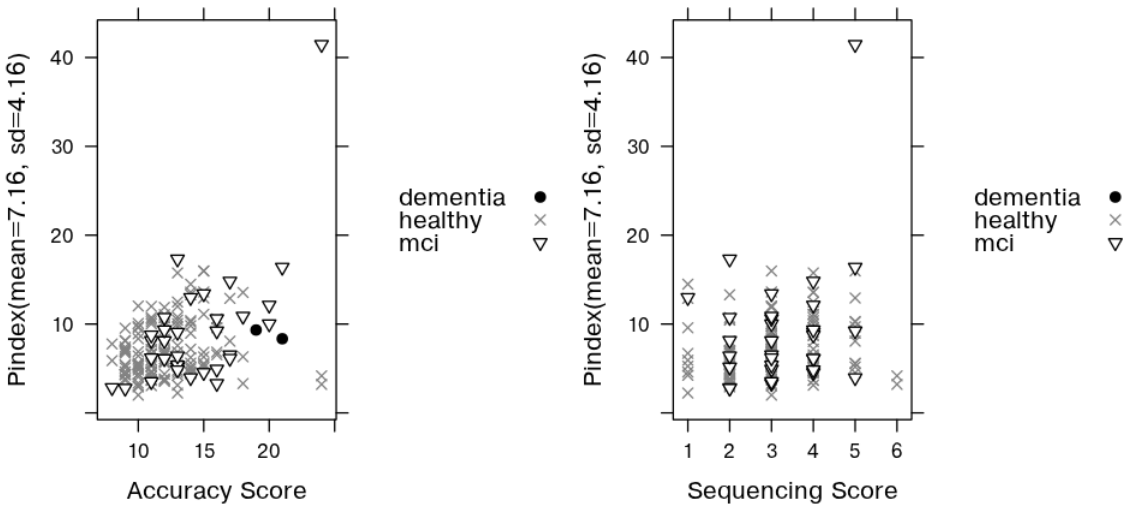


Figure 5.3: DOT subtask order for the participants who completed all eight subtasks. The x-axis represents the subtask sequence position (1..8). The y-axis represents the number of participants. Each bar corresponds to the number of participants that put a particular subtask in the given position of the subtask sequence order.



(a) Scatter plot of Pindex vs. accuracy score

(b) Pindex vs. Sequencing Score

Figure 5.4: Scatter plot of Pindex vs. accuracy score (left) and Pindex vs. sequencing score (right) with participants cognitive diagnosis indicated by point type. The point in the upper right represents a participant who started all DOT subtasks but could only complete two of them.

home DOT features and direct observation scores (accuracy and sequencing scores). The second objective is to study how the correlations between the smart home features and direct observation scores vary as different subsets of participants are considered. We first analyze the relationship for separate participant groups based on how many subtasks they completed then we look at the relationship for the participant groups based on their cognitive diagnosis. The results are summarized in Tables 5.6 and 5.7. In each table, the first row shows the correlation between the entire participant subgroup and the direct observation scores.

Table 5.6: Correlations based on number of subtasks completed.

#Completed subtasks	Sample size (n)	Accuracy score	Sequencing score
2	179	0.79**†	0.45**†
3	174	0.77**†	0.36**†
4	172	0.76**†	0.41**†
5	167	0.75**†	0.37**†
6	154	0.65**†	0.43**†
7	137	0.57**†	0.48**†
8	83	0.43**†	0.49**†

$*p < .05$, $**p < .005$, $\dagger p < 0.05$ with Bonferroni correction for n sample groups

Table 5.7: Correlations based on cognitive diagnosis.

Cognitive Diagnosis	Sample size (n)	Accuracy score	Sequencing score
{D,M,OA,YA}	179	0.79**†	0.45**†
{M,OA,YA}	177	0.80**†	0.43**†
{OA,YA}	145	0.75**†	0.57**†
{YA}	37	0.70**†	0.41**†
{M,YA}	69	0.81**†	0.27*
{M}	32	0.75**†	-0.09
{M,OA}	140	0.78**†	0.34**†

$*p < .05$, $**p < .005$, $\dagger p < 0.05$ with Bonferroni correction for n sample groups

We find that the correlation depends on the heterogeneity in the samples. For example, the strongest correlation is found when examining the population subgroup that contains both MCI and cognitively healthy younger adults and the weakest correlation is found when examining only cognitively healthy individuals. Similarly, we find that the correlation decreases as participant subgroups that completed more subtasks are included. This is because having a large number of incorrect and inefficient tasks helps the learning algorithm to make better predictions due to the variation that

is present in the data. The variations in the samples of cognitively healthy individuals who completed all subtasks are relatively low.

We also find that the correlation is consistently stronger for the accuracy score than the sequencing score. This is because the accuracy score takes into account the mistakes that an individual makes in a subtask while the sequencing score only considers how a participant initiated an activity. When we examine the correlation between the combined set of features and the direct observation scores for the entire population, we see that the coefficient is fairly high ($r = 0.79$ for the accuracy score, $p < 0.005$). This result indicates that automatically derived feature values generated from smart home data do provide valuable information that can be used to assess task quality and that the quality score is fairly consistent with those obtained through direct observation.

Supervised classification of task quality

In this experiment, we train multiple learning models to classify task quality score. We choose the accuracy score as our basis of comparison with automated scores because the correlation coefficients between features derived from sensor data with the accuracy score were consistently higher than the correlation between features from the sensor data and the sequencing score. We divide the scores into two classes using equal-frequency binning. Table 5.8 shows the results of the experiments when all samples are included. All results are generated using leave one out cross valida-

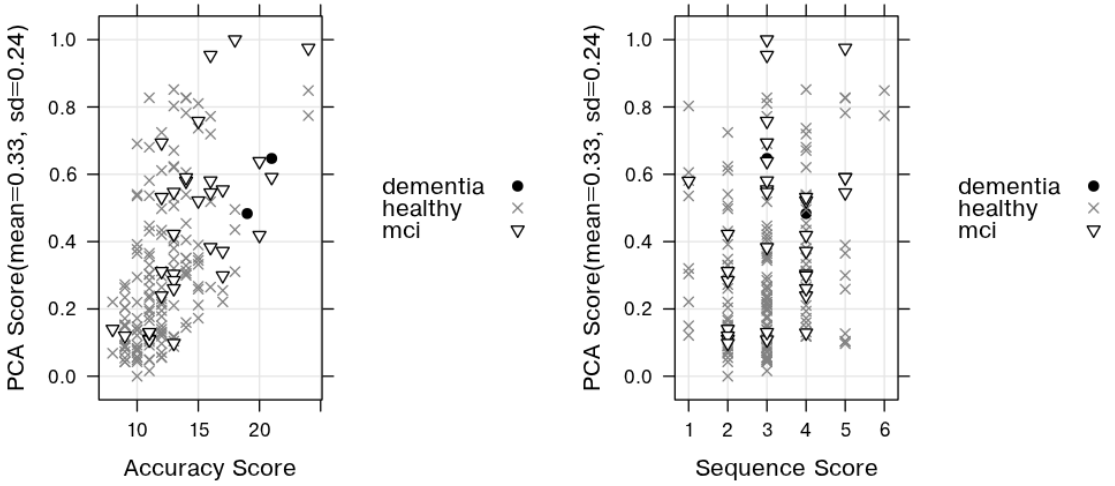
Table 5.8: Performance of the classifiers on the classification of task quality.

Learning algorithm	Accuracy	F-score		AUC
		Class A	Class B	
SVM	80.45	0.84	0.76	0.85
Neural Network	79.33	0.82	0.74	0.85
Naive Bayes	82.13	0.85	0.78	0.88

tion. The machine learning models that are tested include an SMO-based support vector machine, a neural network, and a Naive Bayes classifier. We see that learning algorithms are indeed effective at classifying task quality based on direct observation scores.

5.5.2 Evaluation of unsupervised scoring models

In our next experiment, we analyze the correlation between unsupervised learning model-based generation of a sensor-derived score using Principal Component Analysis and the direction observation-based accuracy score and sequencing score. The objective of this experiment is to test the performance of unsupervised learning models in predicting DOT activity quality scores and determine if the performance of



(a) PCA score vs. Accuracy score

(b) PCA score vs. Sequencing score

Figure 5.5: PCA score vs. Accuracy score (left) and PCA score vs. Sequencing score (right).

Table 5.9: Correlations based on number of subtasks that are completed using PCA. (*$p < 0.05$, **$p < .005$, †$p < 0.05$ with Bonferroni correction for n sample groups.)

#Completed subtasks	Sample size (n)	Accuracy score	Sequencing score
2	179	0.57**†	0.23**†
3	174	0.46**†	0.14
4	172	0.45**†	0.13
5	167	0.50**†	0.13
6	154	0.48**†	0.13
7	137	0.47* †	0.10
8	83	0.43**†	0.10

Table 5.10: Correlations based on cognitive diagnosis computed using PCA. (*$p < .05$, **$p < .005$, †$p < 0.05$ with Bonferroni correction for n sample groups.)

#Cognitive Diagnosis	Sample size (n)	Accuracy score	Sequencing score
{D,M,OA,YA}	179	0.57** †	0.23**†
{M,OA,YA}	177	0.56**†	0.23**†
{OA,YA}	145	0.44**†	0.17
{YA}	37	0.06	0.47**†
{M,YA}	69	0.77**†	0.38*
{M}	32	0.79**†	0.32*
{M,OA}	140	0.51**†	0.17*

an unsupervised algorithm is comparable to that of a supervised learning algorithm. We first analyze the relationship for separate participant groups based on how many subtasks they completed then we look at the relationship for the participant groups based on their cognitive diagnosis. The results are summarized in Tables 5.9 and 5.10. Figure 5.5 shows the plot of the PCA score that is obtained by reducing the feature space to a single dimension as a function of the accuracy and sequencing scores.

Similar to previous observations, we find that the correlation depends on the heterogeneity in the samples. For example, the strongest correlation is found when examining the population subgroup that contains both MCI and cognitively healthy younger adults. The correlation coefficient between the unsupervised score and the direct observation accuracy score is 0.57 $(p < 0.005)$. This indicates that a fairly strong positive correlation exists between the automated scores and experimenter-generated scores of task quality. Furthermore, this value is similar to the values generated for the SVM model, which indicates that task quality can be computed directly using smart home sensor data without relying on training from human-provided scores.

5.5.3 Evaluation of cognitive assessment models

The second goal is to design a learning approach to automate cognitive health assessment based on smart home features. For this study, we map each participant to one of three labels: CH, MCI, or Dementia (D). We use labels provided by clinical testing to train the learning algorithm. Note that this data is based on a battery of standardized and experimental neuropsychology tests administered in a laboratory setting and not on the smart home data. We handle the assessment as a set of binary classification problems.

Class imbalance is a challenge in learning a discriminative model between these three classes. While there are 145 cognitively healthy individuals, there are 32 individuals with MCI and only 2 participants with dementia. Part of this imbalance is because many dementia participants had difficulty completing basic everyday tasks independently. Class imbalance affects classification performance because machine learning models tend to label the points with the majority class label. To address this issue, we use cost sensitive versions of machine learning algorithms for each of the base classifiers. A cost sensitive classifier assigns misclassification costs separately for individual class labels and reweights the samples during training according to this cost. This allows the classifier to achieve overall strong performance even when the training points are not evenly divided among the alternative classes [145], as is the

case with this dataset.

We initially train a learning algorithm to label CH and MCI participants. We use PCA to reduce the dimensionality of the feature vector and train a cost-sensitive version of a support vector machine. We compare this with an alternative approach in which we handled the class imbalance by under-sampling the majority class so that the ratio of the Cognitively Healthy group to the MCI group is $2 : 1$. The results of this experiment are summarized in Table 5.11. To compare automated diagnosis based on smart home features with diagnosis based on direction observation features, we train a learning algorithm to map direct observation scores to cognitive health diagnosis labels. The AUC value for this mapping is 0.68 in the best case (using nave Bayes and under sampling). The predictive performance overall is not as strong as we would like to see for this case, in part because performance of CH and MCI participants is actually quite similar on familiar activities such as those used in the DOT. The individuals in these two groups do have quite a bit of overlap in functional performance as is evident in Figure 5.4.

Our next objective is to compare the Cognitively Healthy group with the Dementia group. We have a limited number of data points for the dementia group because out of 16 dementia participants only 2 completed the DOT. Hence, we perform an exploratory experiment to compare these two groups by under-sampling the Cognitively Healthy class so that the ratio of Cognitively Healthy data points to De-

mentia data points is 4 : 1 and ensuring that one Dementia participant would be used each time for training and the other would be used for testing. The results are averaged and summarized in Table 5.12. As expected, these two groups are much easier to distinguish. To obtain stronger classification performance, we can include all participants with dementia and represent the sensor features as missing for participants with dementia and noting the number of tasks that were completed as 0. These experiments provide evidence that the learning algorithm can indicate the cognitive health of an individual based on activity performance.

Table 5.11: Performance of the machine learning classifiers on the supervised classification of cognitive health (MCI/ Cognitively healthy).

Learning algorithm	F-score		AUC
	Class A	Class B	
PCA + SVM with Cost Sensitive Learning	0.39	0.73	0.64
Under sampling of Majority Class + Bagged SVM	0.34	0.44	0.61

5.6 Disscussion and observations

Researchers have hoped that ubiquitous computing technologies could be used to support health monitoring and aging in place. This study provides an indication

Table 5.12: Performance of the machine learning classifiers on the supervised classification of cognitive health (Dementia/Cognitively healthy).

Learning algorithm	F-score		AUC
	Class A	Class B	
Under sampling + Bagged SVM	0.52	0.52	0.58
Missing Values+ SVM	0.93	0.99	0.94

that with smart home sensor data and machine learning algorithms it is possible to automatically predict the quality of daily activities.

One must carefully interpret the results that we have mentioned. We note that the correlation (r) between smart home features and task accuracy scores is statistically significant. We can conservatively analyze the correlation coefficient using a coefficient of determination. We square the correlation coefficient to obtain the coefficient of determination. A coefficient of determination of 0.62 ($r = 0.79$) means that the 62% of the variation in the dependent variable can be explained by the variation in the independent variable. Our current results show that our method explains nearly 62% variations in the direct observational scores. Unexplained variation can be attributed to limitations of sensor system infrastructures and algorithms.

This implies that smart home technologies provides valuable information to

assess the quality of daily activities. On the other hand, predicting cognitive health based on the performance on activities of daily living is an active research area in clinical research [130]. Thus, we believe that smart home based technologies can monitor activities of daily living and predict the cognitive health of an individual. Our results indicate this as a possibility.

We observe from the experiments that the performance of automatic task quality prediction depends on the type of training samples. The learning algorithm offers accurate predictions when the training samples contain heterogeneous data points of both well-conducted activities and poorly-conducted activities. We observe that sequencing features are less indicative when all of the participant samples are cognitively healthy, while parallel features are indicative when we include MCI and younger adult participant samples. We therefore conclude that researchers need to carefully define and extract appropriate features from sensor data to use in building an assessment model. In addition, for our study the baseline for performance is a direct observation score based on coders observation of task performance. Two coders independently assigned scores to participants based on specific criteria as they directly viewed the participant's task performance. We cannot ignore that there may be some error or bias in these direct observation scores. This error can be mitigated by increasing the number of clinicians scoring the activities or by automatically detecting and correcting for bias.

Our approach to perform automated cognitive health assessment using smart home sensors and algorithms has a few limitations. The first limitation is due to the coarse granularity of the home-based sensors. While environment sensors face fewer practical issues of user acceptance, placement, and battery charge, our algorithms would benefit from data provided by wearable, smart phone, and object sensors. Also, many participants with cognitive difficulties were not able to complete the activities. We can address this issue by increasing our sample of participants. We note that the complexity of DOT was necessary to capture differences in task performance between cognitively healthy and MCI participants, but additional tasks that are less complex but still involve multi-tasking can be devised for future studies.

Similarly, the limitations of experimental methodology are that assessment technique relies on participants completing scripted activities in a single smart home setting. These types of methods are argued to be ecologically valid [24] but participants can perform activities in an unnatural manner due to the unfamiliar environment, the scripted manner of the activity, or the awareness of being monitored. In addition, we use direct observation scores and clinician-based cognitive diagnosis as ground truth labels to train our learning models. Instead, we would like to learn models based on differences in natural activity performance between individuals who are known to be cognitive healthy and those who are known to have cognitive difficulties. Finally, some of the derived features rely on human annotation of sensor data. We can

avoid this annotation step by using activity recognition algorithms that can recognize interleaved and parallel activities as well as activity steps.

We showed that machine learning algorithms can be designed to perform automated assessment of task quality based on smart home sensor data that is collected during task performance. Our results indicate that smart homes and ubiquitous computing technologies can be useful for monitoring complex everyday functions and to automate assessment of daily activities. This capability is valuable for monitoring the well-being of individuals in their own environments.

In this and the previous chapter, we presented two cross-sectional studies in which participants perform eight different activities and a complex real lifelike activity in our smart home testbed. We also introduced machine learning methods to model the quality of these simple activities as well as the complex activity and use that model to predict the cognitive health of the participants. The sensor data in these two cross-sectional studies are collected when volunteer participants performed pre-defined set of activities in our smart home testbed. In contrast to these two chapters, the next two chapters discuss a smart home-based longitudinal study. In this longitudinal study, sensor data is collected from 18 single resident smart home apartments for over 2 years without manipulating and altering residents routines. We present algorithms to model the everyday behavior of a smart home resident and predict their cognitive and physical health utilizing the longitudinal sensor data.

CHAPTER 6. LONGITUDINAL ANALYSIS OF SMART HOME-BASED BEHAVIOR DATA

In this chapter, we discuss a longitudinal smart home study in which we collect sensor data from 18 single resident smart homes for over 2 years. We propose the Clinical Assessment Using Activity Behavior (CAAB) algorithm to model everyday behavior of a smart home resident and use machine learning algorithm to predict the cognitive and physical health of a smart home resident utilizing this longitudinal sensor data.

6.1 Background

We investigate whether smart home-based behavior data can be used to predict an individual's standard clinical assessment scores. We hypothesize that a relationship does exist between a person's cognitive/physical health and his/her daily behavior as monitored by a smart home. We monitor the daily behavior of a resident using smart home sensors and quantify their cognitive/physical health status using standard clinical assessments. To validate this hypothesis, we develop an approach to predict the cognitive and physical health assessment scores by making use of real-world smart

home sensor data.

We propose a Clinical Assessment using Activity Behavior (CAAB) approach to predict the cognitive and mobility scores of smart home residents by monitoring a set of basic and instrumental activities of daily living. CAAB first processes the activity-labeled sensor dataset to extract activity performance features. CAAB then extracts statistical activity features from the activity performance features to train machine learning algorithms that predict the cognitive and mobility scores. To evaluate the performance of CAAB, we utilize sensor data collected from 18 real-world smart homes with older adult residents. An activity recognition (AR) algorithm labels collected raw sensor data with the corresponding activities.

CAAB utilizes sensor data collected from actual smart homes without altering the resident's routine and environment. Therefore, the algorithmic approach offers an ecologically valid method to characterize the ADL parameters and assess the cognitive and physical health of a smart home resident [24]. To the best of our knowledge, our work represents one of the first reported efforts to utilize automatically-recognized ADL parameters from real-world smart home data to predict the cognitive and physical health assessment scores of a smart home resident.

Table 6.1: Major notations and meanings in CAAB

n	Number of activities
T	Total number of data collection days
A	Set of n activities being modeled
P_i	Activity performance feature vector for activity i modeled over data collection period T
$P_{i,d,t}$	Activity performance feature d for activity i activity on day t
j	Time point at which clinical measurements are made
S_j	Clinical assessment score measured at time point j
W	Sliding window size

6.2 Problem formulation

We assume that smart home sensors produce a continuous sequence of time-stamped sensor readings, or sensor events. These sensors continuously generate raw sensor events while residents perform their routine activities of daily living. We use an activity recognition algorithm to automatically annotate each of these sensor events with a corresponding activity label. Activity recognition algorithms map a sequence of raw sensor events onto an activity label A_i, where the label is drawn from the predefined set of activities $A = \{A_1, A_2, \ldots, A_n\}$. Our activity recognition algorithm generates a label that corresponds to the last event in the sequence (i.e., the label indicates the activity that was performed when the last event was generated). Activities from set A can be recognized even when the resident interweaves them or multiple residents perform activities in parallel.

CAAB extracts activity performance features from activity-labeled smart home sensor data and utilizes these features to predict standard clinical assessment scores. Therefore, there are two steps involved in CAAB:

- Modeling the ADL performance from the activity-labeled smart home sensor data.

- Predicting the cognitive and mobility scores using a learning algorithm.

Activity modeling: We extract a d-dimensional activity performance feature

vector $P_i =< P_{i,1}, \ldots, P_{i,d} >$ to model the daily activity performance of an activity A_i. Observation $P_{i,d,t}$ provides a value for feature d of activity A_i observed on day $t \quad (1 \leq t \leq T)$. The set of all observations in P_i is used to model the performance of A_i during an entire data collection period between day 1 and day T.

Additionally, during the same data collection period, standard clinical tests are administered for the resident every m time units, resulting in clinical assessment scores $S_1, S_2, \ldots, S_p$ $(p = T/m)$. In our setting, the clinical tests are administered biannually ($m = 180$ days). Therefore, the clinical measurements are very sparse as compared to the sensor observations. The baseline clinical measurement, S_1, is collected after an initial 180 days of smart home monitoring.

Clinical assessment/ Clinical assessment scores prediction: CAAB's goal is to accurately predict clinical assessment scores at time k, or S_k, using activity performance data P_i between time points j and k, $j < k$.

CAAB relies on an activity recognition (AR) algorithm to generate labeled data for the performance feature vector that is an integral component of activity modeling. The method for activity recognition is explained in Section 2.1.6 and explored in detail elsewhere [32]. Here, we utilize our own AR algorithm and focus on the additional steps that comprise CAAB.

6.3 Experimental setup

We use CAAB approach to analyze data collected in our CASAS smart homes[2] [29] and in our corresponding clinical measurements. Below, we explain the smart home testbed, smart home sensor data, and standard clinical data that are collected as a part of the study.

6.3.1 CASAS Smart home testbed

The CASAS smart home testbeds used in this study are single-resident apartments, each with at least one bedroom, a kitchen, a dining area, and at least one bathroom. Refer to Section 2.1.3 for more details on CASAS longitudinal smart home testbed as well as the drawings in Appendix C.

The residents perform their normal activities in their smart apartments, unobstructed by the smart home instrumentation. Each sensor event is represented by four fields: date, time, sensor identifier, and sensor value. The raw sensor data does not contain activity labels. We use our AR activity recognition algorithm, described in Section 2.1.6, to label individual sensor events with corresponding activity labels

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

as shown in Figure 2.7.

6.3.2 Residents

Residents included 18 community-dwelling seniors (5 females, 13 males) from a retirement community. All participants are 73 years of age or older (M = 84.71, SD = 5.24, range $73-92$) and have a mean education level of 17.52 years (SD = 2.15, range $12-20$). At baseline S_1, participants were classified as either cognitively healthy ($N = 7$), at risk for cognitive difficulties ($N = 6$) or experiencing cognitively difficulties ($N = 5$). One participant in the cognitively compromised group met the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR) criteria for dementia [10], while the other four individuals met criteria for mild cognitive impairment (MCI) as outlined by the National Institute on Aging-Alzheimer's Association work group [5]. Participants in the risk group had data suggestive of lowered performance on one or more cognitive tests (relative to an estimate of premorbid abilities), along with sensory and/or mobility difficulties.

6.3.3 Clinical tests

Clinicians *biannually* administered standardized clinical, cognitive, and motor tests to the residents. The tests included the Timed Up and Go mobility measure

(TUG) as well as the Repeatable Battery for the Assessment of Neuropsychological Status measure of cognitive status (RBANS) as detailed in Table 2.3. We create a clinical dataset using TUG and RBANS scores obtained from biannual clinical tests. Figure 6.1 plots the distribution of these two scores against the ages of the participants.

6.4 Modeling activities and mobility

6.4.1 Modeling performances of activities and mobility performances

The first CAAB step is to model the performance of the activities in set A. We model activity performance by extracting relevant features from the activity-labeled sensor data. For each activity $A_i \in A$, we can represent such performance features using the d-dimensional activity performance feature vector $P_i =< P_{i,1}, P_{i,2}, \ldots, P_{i,d} >$.

Depending upon the nature of the sensor data and the performance window we want to monitor, we can aggregate activity performance P_i for activity A_i over a day, week, or other time period. In our experiments, we aggregate activity performance features over a day period (the time unit is one day). For example, if we calculate the sleep activity performance $P_{i,1,t}$ as the time spent sleeping in the bedroom on day t, the observation $P_{i,1,t+1}$ occurs one day after observation $P_{i,1,t}$. For each individual, we calculate activity performance features for the entire data collection period T for

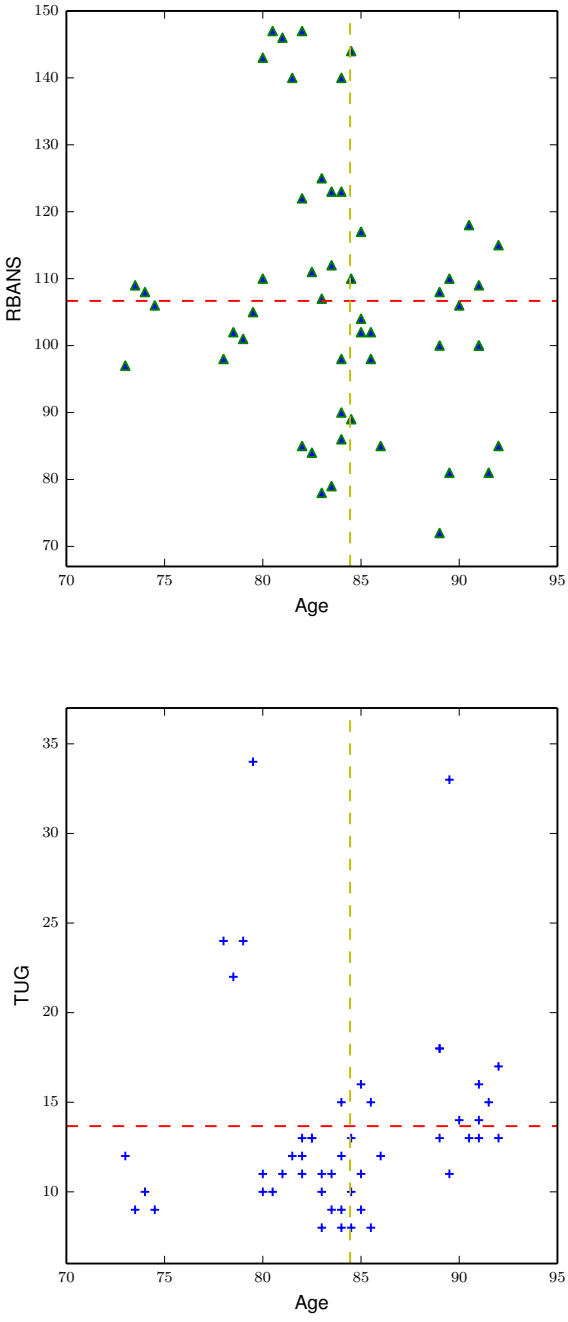


Figure 6.1: Distribution of RBANS (top) and TUG (down) clinical assessment scores in the y-axis with respect to age in x-axis. The horizontal line represents a mean clinical score and the vertical line represents the mean age.

Table 6.2: Activity performance features extracted from the activity-labeled smart home sensor data.

Group	Variable	Features
Mobility	Mobility	Total distance traveled, #Total sensor events
Sleep	Sleep	Sleep duration, #Sleep sensor events
	Bed toilet transition	Bed toilet transition duration
ADL	Cook	Cook duration
	Eat	Eat duration
	Relax	Relax duration
	Personal hygiene	Personal hygiene duration
	Leave home	Leave home duration

all activities in the activity set $A \quad (1 \leq t \leq T)$.

For our experiments, we model activity performance using two ($d = 2$) specific activity performance features, a time-based feature and a sensor-based feature $\{P_{i,1}, P_{i,2}\}$. Feature $P_{i,1}$ represents the duration of activity A_i and $P_{i,2}$ represents the number of sensor events generated during activity A_i. We have provided evidence in previous studies that these two features are generalizable to other activities, are easily interpretable, and can model how the residents perform their daily activities [40]. In addition to capturing activity performance, we also represent and monitor a person's overall mobility. Mobility refers to movement generated while performing varied activities (as opposed to representing a single activity of its own) and is therefore represented using two different types of features: the number of sensor events triggered throughout the home and the total distance that is covered by movement throughout the course of a single day (see Table 6.2).

6.4.2 Selection of ADLs

In this study, we model a subset of automatically-labeled resident daily activities. These activities are sleep, bed to toilet (a common type of sleep interruption), cook, eat, relax, and personal hygiene. We also capture and model a resident's total mobility in the home.

Sleep

The effects of aging include changes in sleep patterns that may influence cognitive and functional status. For example, individuals over the age of 75 have been found to experience greater fragmentation in nighttime sleep (e.g., [98, 103]), which concurrently causes decreased total sleep time and sleep efficiency [62]. Sleep problems in older adults can affect cognitive abilities [69, 134] and have been associated with decreased functional status and quality of life [49, 90]. Moreover, individuals with dementia often experience significant disruption of the sleep-wake cycle [51]. Thus, the effects of sleep on the health of older adults are important clinical construct that both clinicians and caregivers are interested in understanding [43, 135].

Using AR, we recognize sensor events that correspond to sleep (in the bedroom, as opposed to naps taken outside the bedroom) and bed-to-toilet activities. We then extract the time spent and number of sensor events features that correspond to these two activities. As listed in Table 6.2, four features model a smart home resident's sleep activity. The value for the time-based sleep feature is calculated as the total number of minutes spent in sleep on a particular day and the value for the sensor-based sleep feature is calculated as the number of sensor events that are triggered over the course of one day while the resident slept. Similarly, the time-based bed to toilet feature is calculated as the total number of minutes spent in bed to toilet activity on a particular day. We exclude the sensor-based feature that calculate number of

times sensor events are triggered on bed to toilet activity because our data shows that the number of sensor events generated when performing the bed to toilet activity is often very low. Because of the known importance of sleep and its relationship with physical and cognitive health, we conduct a separate analysis of sleep and bed to toilet parameters from the other activities that are analyzed as a group [43, 90].

Mobility

Mobility is the ability of an individual to move around their home environment and the community. Mobility impairments limit an individual's ability to maintain independence and quality of life and are common predictors of institutionalization among older adults [65]. Evidence supports a close connection between executive brain function and walking speed [17, 129]. Prior studies have also demonstrated a relationship between mobility, gait disorders, cognitive decline, and the risk of disability [35, 60, 144]. Therefore, we separately model mobility as an everyday behavioral feature. We model the mobility of a smart home resident based on the number of sensor events they trigger and the total distance they cover in a day while in the home (estimated based on known distances between motion sensors placed in the home). As listed in Table 6.2, the value for the distance-based mobility feature is calculated as the total distance covered by a resident in one day (our aggregation time period) while inside the home. Similarly, the value for the sensor-based mobility feature is calculated as the number of sensor events that a resident triggers over the

course of one day while moving around in the home.

Activities of Daily Living

Basic activities of daily living (e.g., eating, grooming) and the more complex instrumental activities of daily living (IADLs; e.g., cooking, managing finances), are fundamental to independent living. Data indicate that increased difficulties in everyday activity completion (e.g., greater task inefficiencies, longer activity completion times) occur with older age [94, 131]. Clinical studies have also demonstrated that individuals diagnosed with MCI experience greater difficulties (e.g., increased omission errors) completing everyday activities when compared with healthy controls [9, 53, 110, 133]. Furthermore, individuals with severe cognitive problems, such as AD, have difficulty in both initiating and completing basic activities [59]. While research has shown that IADLs are affected earlier in dementia than basic self-care tasks [110], a detailed understanding of the course of functional change between normal aging and dementia is still a gap in the research [53, 113, 147]. Therefore, clinicians argue the importance of understanding the course of functional change given the potential implications for developing methods for both prevention and early intervention [130, 131].

In our work, we consider five activities of daily living (in addition to sleep): cook, eat, personal hygiene, leave home, and relax. We note that the "relax" activity represents a combination of watching TV, reading, and napping that typically takes

place in a single location other than the bedroom where the resident spends time doing these activities, such as a favorite chair. We focus on these activities because they are activities of daily living that are important for characterizing daily routines and assessing functional independence. For each of these activities, we calculate the total activity duration. Our data shows the number of sensor events generated when performing these activities is often very low. Thus, for these activities, we exclude features that calculate number of times sensor events are triggered. As listed in Table 6.2, we calculate the value for the time-based ADL feature as the total number of minutes spent in an activity on a particular day.

6.4.3 Activity feature extraction

The second CAAB step is to extract statistical features from the activity performance vector. CAAB extracts features from the time series-based representation of activity performance and uses these to train a machine learning algorithm. Namely, we extract four standard time series features and one new change feature. We will refer to these five features as statistical activity features. Table 6.3 lists the complete set of activity features.

Table 6.3: Statistical activity features (μ is the mean of the activity performance features p of size n).

Id	Statistical features	Definition	Formula
1	Variance	Variance is the measure of spread.	$\text{Var(p)} = \sum_{k=1}^{n}(p_i - \mu)^2$
2	Autocorrelation	Autocorrelation(AC) is the similarity between observations that are displaced in time. We calculate autocorrelation at lag 1.	$\text{AC-lag1}(p) = \frac{\sum_{i=1}^{n-1}(p_i-\mu)(p_{i+1}-\mu)}{\sum_{n=1}^{n}(p_i-\mu)^2}$
3	Skewness	Skewness measures the degree of asymmetry in the distribution of values.	$\text{skewness(p)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(p_i-\mu)^3}{(\frac{1}{n}\sum_{i=1}^{n}(p_i-\mu)^2)^{3/2}}$
4	Kurtosis	Kurtosis measures the amount of peakedness of the distribution toward the mean.	$\text{kurtosis(p)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(p_i-\mu)^4}{(\frac{1}{n}\sum_{i=1}^{n}(p_i-\mu)^2)^3}$
5	Change	Change characterizes the amount of change in an individual's activity performance over time.	Algorithm 2

Statistical activity features

To calculate the first four features, CAAB runs a sliding window (e.g., window size, $W = 30$ days) over each of the activity performance features listed in Table 6.2 and calculates variance, autocorrelation, skewness, and kurtosis using the observations from data that falls within the sliding window. The sliding window starts at one clinical assessment time point and ends at the next assessment time point, thus capturing all of the behavior data that occurred between two subsequent assessments. For example, CAAB calculates the variance, autocorrelation, skewness, and kurtosis of the duration feature for each activity based on duration observations that fall inside each W-sized data window. CAAB repeats the process and calculates these four statistical activity features for all other activity performance features for all of the activities in set A.

Before calculating these features, CAAB first removes the time series trend from the sliding window observations in order to remove the effect of non-stationary components (e.g. periodic components) in the time series [37]. For this step, CAAB fits a Gaussian or a linear trend to the data within the sliding window. CAAB then detrends the data by subtracting the fitted trend from the data. CAAB slides the window by one day (skip size=1) and re-computes all of the statistical activity features. For each feature, CAAB slides a window through the sensor home data and computes the final feature values as an average over all of the windows. Algorithm 1

Algorithm 1 CAAB approach

Input: Activity performance features

Output: Statistical activity features

Initialize: Feature vector

// T_1 and T_2 are two consecutive clinical testing time points

Given: T_1, T_2

Given: skip size = 1

while $T_1 < (T_2 - W)$ **do**

    for each activity performance feature **do**:

        Place a window of size W at T_1.

        Remove missing observations and detrend based on the observations that fall into this window.

        Calculate the variance, autocorrelation, skewness, kurtosis and change features (Algorithm 2) using the observations in the window.

        Append these values to the feature vector.

        $T_2 = T_1$ + skip size

    end foreach

end while

return average(Feature matrix)

explains the steps.

In addition to these standard four different time series features, we propose a fifth feature, a change-based feature, to characterize the amount of change in an individual's activity performance. Algorithm 2 details the steps in calculating this new feature. In order to compute this feature, CAAB uses a sliding window of size W days and divides an activity performance feature observations that fall in W into two different groups. The first group contains feature observations that fall in the first half of W and second group contains feature observations that fall in the other half. CAAB then compares between these two groups of feature observations using a change detection algorithm. For the current work, we use the Hotelling-T test algorithm [66]. However, we can also apply other change detection algorithms. CAAB then slides the window by one day (skip size = 1) and re-computes the change feature. CAAB calculates the final change value as the average over all windows. Similar to the other four statistical activity features computed in the previous section, CAAB computes the value of the change feature for each of the activity performance features listed in Table 6.2.

We note that the change feature is different from the variance feature that CAAB calculates earlier. While variance measures the variability of samples around its mean, the change feature empirically calculates the "chance" of observing a change when two sample groups each of size n from the given activity performance features

are compared with each other. Here, a higher amount of detected change indicates a greater chance of detecting changes in the activity performance feature.

6.4.4 Clinical assessment

In the final step, CAAB predicts the clinical assessment scores of the smart home residents using the activity performance features computed from the activity-labeled sensor data. CAAB first aligns the sensor-based data collection date with the clinical assessment-based data collection date before extracting statistical activity features. After extracting features and aligning the data, CAAB then trains a supervised machine learning algorithm and predicts the clinical assessment scores.

To accomplish this goal, CAAB extracts statistical activity features from the activity performance features that lie between any given two consecutive clinical testing points, t_1 and t_2. Similarly, it obtains the clinical score S_2 (or S_1) at time point t_2 (or t_1) . We consider the pair, statistical activity features and clinical score S_2, as a point in the dataset and repeat the process for all of the smart home residents and for every pair of the consecutive clinical testing points. Algorithm 3 summarizes the steps involved to prepare the dataset.

The final step in the CAAB is to predict the clinical assessment scores. CAAB trains a learning algorithm to learn a relationship between statistical activity features

Algorithm 2 Calculation of change feature

Input: Activity performance features
Initialize: CH = [ ]
// T_1 and T_2 are two consecutive clinical testing time points
Given: T_1,T_2
Given: skip size = 1
W = window size
while do $T_1 < (T_2 - W)$:
    for each activity performance feature **do**:
        Place window of size W at T_1.
        Remove missing values that fall into this window.
        Put first half of W in the group A and second half in the group B.
        // Returns True or False.
        change = Hotelling T-test (A,B)
        append(CH, change)
        $T_1 = T_1$ + skip size
    end foreach
end while
return average(CH)

and the clinical assessment scores using the dataset that is constructed. In this step, for each resident, at each time point (except the first one), CAAB predicts the clinical assessment scores using a learning algorithm.

We note that CAAB predicts clinical assessment scores based on the relationship that the learning algorithm models between the clinical assessment scores and behavior features. We followed this approach because there are very few clinical observations for a resident. Furthermore, we note that CAAB computes activity performance features by temporally following an individual over a period and computes statistical activity features by comparing past observations with current observations. In this way, CAAB uses an individual as their own baseline for predictive assessment.

6.5 Experimental evaluation

6.5.1 Dataset

As explained in Section 6.3.1, the CASAS middleware collects sensor data while monitoring the daily behavior of 18 smart home senior residents for approximately 2 years. We use the AR activity recognition algorithm to automatically label the sensor events with the corresponding activity labels. By running CAAB on the (activity-labeled) sensor data, we compute activity performance features and extract activity features from them. CAAB then creates a training set by combining the activity

Algorithm 3 Training set creation

1: Output: Training set to train the learning algorithm

2: Input: Activity performance features for all residents

3: Initialize: Empty training set TrSet

4: **for each** resident **do**

5: **for each** consecutive clinical testing point T_1 and T_2 **do**

6: F = CAAB (activity performance features between T_1 and T_2)

7: S = clinical score(T_1, T_2)

8: Append(F,S,TrSet)

9: **end foreach**

10: **end foreach**

Table 6.4: Details of the training set used in CAAB approach.

	#Instances	#Features
RBANS	52	50
TUG	50	50

features and the corresponding clinical assessment scores (RBANS and TUG) to train a learning algorithm. Table 6.4 provides details of the final training dataset.

6.5.2 Prediction

We perform the following four different prediction-based experiments to evaluate the performance of CAAB approach and its components : 1) We first evaluate the overall CAAB performance in predicting clinical assessment scores. Here, we train CAAB using the complete set of available features. We compare results from several representative supervised learning algorithms. 2) We then investigate the importance of different activity feature subsets by observing the resulting performance of CAAB in predicting the clinical assessment scores. 3) Next, we investigate the influence of parameter choices on performance by varying CAAB parameter values and analyzing the impact on prediction performance. 4) In the final experiment, we

compare CAAB performance utilizing AR-labeled activities with a baseline method that utilizes random activity labels.

We evaluate all of the above experiments using linear correlation coefficient (r) and mean squared error (RMSE). All performance values are generated using leave-one-out cross validation. The data for each participant is used for training or held out for testing, but is not used for both to avoid biasing the model. We use the following methods to compute our performance measures.

- Correlation coefficient(r): The correlation coefficient between two continuous variables X and Y is given as: $r_{X,Y} = \frac{cov(X,Y)}{\sigma_x \sigma_y}$ where σ_x and σ_y are the standard deviations of X and Y and $cov(X,Y)$ is the covariance between X and Y. In our experiments, we evaluate the correlation between the learned behavior model and clinical assessment scores. We will interpret the experimental results based on the absolute value of the correlation coefficient because our learning algorithm finds a non-linear relationship between statistical activity features and the clinical assessment scores.

- Root Mean Squared Error (RMSE): If $\hat{y}$ is a size-n vector of predictions and y is the vector of true values, the RMSE of the predictor is RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$

Overall CAAB prediction performance

To validate the overall performance of CAAB performance, we compute correlations between the CAAB-predicted clinical assessment scores and the provided clinical assessment scores using the complete set of activity features and three different supervised learning algorithms:

- Support Vector Regression (SVR): Support vector regression uses support vector machine algorithm to make numeric predictions. The learning model can be expressed in term of support vectors and kernel functions can be used to learn a non-linear function. SVR uses the epsilon insensitive loss function that ignores errors that are smaller than threshold $\epsilon > 0$. We use a linear kernel to generate all our prediction-based performance results [164].

- Linear Regression (LR): Linear regression models the relationship between the class and the features as the weighted linear combination of the features. The weights are calculated from the training data often using the least square approach.

- Random Forest (RF): Random forest builds an ensemble learner by creating multiple decision trees on different bootstrap samples of the dataset. It averages the predictions from these decision trees to make the prediction [164].

Table 6.5: Overall prediction performance of the different learning algorithms. (*$p <$ 0.05, **$p < 0.005$)

Score Type	Measure	SVR	LR	RF
RBANS	r	0.72**	0.64**	0.52**
	RMSE	14.90	20.25	13.66
TUG	r	0.45**	0.41*	0.41**
	RMSE	5.87	7.62	5.22

As listed in Table 6.5, we observe that the performances of the learning algorithms in predicting the clinical assessment scores are similar. We also observe that the correlation values are all statistically significant. Because SVR performed best overall, we will conduct all of the remaining experiments using this approach. Additionally, we observe that the overall correlation between the predicted TUG scores and the actual TUG scores are weaker than the predicted RBANS and actual RBANS scores. The weaker correlation is likely due to the fact that there are only two activity performance features (mobility and leave home) that represent the mobility of an individual. Other activities such as cook, bed to toilet, and relax do not adequately represent the mobility of a resident.

CAAB prediction performance based on activity feature subsets

We perform a second set of prediction-based experiments using different subsets of statistical activity features to study and find the important sets of features as listed as follows:

1. We evaluate the prediction performances of the learning algorithm when it is trained using different subsets of statistical activity features.

2. We evaluate the result of using statistical activity features that belong to various subsets of ADLs.

In the first experiment, we study the significance of five major types of statistical activity features (autocorrelation, skewness, kurtosis, variance, and change) that CAAB extracts from the activity performance features. To perform this experiment, we create five different training sets, each of which contains a subset of the statistical activity features. For example, the first training set contains all of the variance-based features; the second training set contains all of the autocorrelation-based features etc. Using these training sets, we train five separate support vector machines. As listed in Table 6.6, we note that the performance of the SVR in predicting clinical assessment scores using the variance of the activity features is strong as compared to other major types of statistical activity features. Therefore, we hypothesize that the variance of activity performance is an important predictor. Additionally, we observe

that skewness-based feature is important for predicting TUG clinical scores while it was slightly weaker for RBANS predictions.

For the second CAAB feature-based experiment, we study the relationship between the clinical assessment scores and the statistical activity features subsets that belong to various groups of ADLs. We create nine different ADL groups, each of which contains a combination of one or more activities (out of seven activities) and/or mobility. For each combination, we create a training set containing all statistical activity features belonging to the activities in that combination. In total, we create nine different training sets. As listed in Table 6.7, we make the following three observations:

1. In terms of single variables, sleep had the highest correlation with RBANS ($r = 0.51$). In contrast, mobility showed little correlation with either clinical score.

2. We observe that correlation is higher when we combine variables. Specifically, including automatically-recognized ADLs improved the correlation further for both RBANS ($r = 0.61$) and TUG ($r = 0.48$). RBANS showed highest correlation when all features are used ($r = 0.72$). We also note that the mobility-based features contain information from raw sensor events and do not include activity information and has a weak correlation with little correlation with either clinical score.

3. In the case of TUG, the only two variable combinations that lacked a significant correlation included mobility. Once again, adding automatically-recognized activities generally increases the correlation.

These results show that a relationship exists between RBANS and TUG clinical assessment scores with combined smart home-based parameters of sleep and ADLs. Our observations are interesting and align with results from prior clinical studies that have found relationships between sleep and ADL performance with cognitive and physical health [69, 111]. Furthermore, we also note that our observations are computed by making use of automated smart home sensor data and actual clinical assessment scores. The smart home sensor data are ecologically valid because the smart home collects data from the real world environment and CAAB extracts features without governing, changing, or manipulating the individual's daily routines.

CAAB performance using different parameters

We perform two different experiments to study the effect of parameter choices on CAAB. In these two experiments, we train the learning algorithm using the complete set of features. We first study how the activity features extracted at different window sizes will affect the final performances of the learning algorithm. Second, we repeat the steps of the first experiment to study the effect of using different trend removal techniques.

In the first experiment, we compare performance using different window sizes

Table 6.6: Correlation coefficient (r) and RMSE values between SVR predicted RBANS and TUG scores when SVR is trained using different types of statistical activity features (*$p < 0.05$, **$p < 0.005$).

Score Type	Measure	Change	ACF	Skewness	Kurtosis	Variance	All Features
RBANS	r	0.29	0.17	0.30*	0.21	0.49**	0.72**
	RMSE	25.77	21.39	19.90	25.19	17.76	14.94
TUG	r	0.06	0.05	0.43**	0.06	0.31*	0.45*
	RMSE	6.05	6.12	5.23	6.60	5.56	5.87

and the SVR learning algorithm. We summarize the results in Figure 6.2. We observe that the strength of the correlation between the actual clinical assessment scores and predicted scores using features derived from smaller and mid-sized window is stronger than the larger-sized windows. One possible explanation is that larger windows encapsulate more behavior trends and day-to-day performance variation may be lost. Therefore, we use mid-sized (30 for RBANS and 55 for TUG) windows for all of our experiments.

In the second experiment, we compare three different trend removal techniques. We create three different training sets that result from removing a Gaussian trend, a linear trend, and no trend removal. The results are showed in Figure 6.2. We observe

Table 6.7: Correlation coefficient (r) and RMSE values between SVR-predicted RBANS and TUG scores when the SVR is trained using features from different activities (*$p < 0.05$, **$p < 0.005$).

Score Type	Measure	Sleep	Mobility	ADL	Mobility+ Leave Home	ADL+ Leave home	Sleep+ Mobility	Sleep+ ADL	Sleep+ ADL+ Leave home	Mobility+ ADL	All Features
RBANS	r	0.51**	0.08	0.35*	0.18	0.27	0.41*	0.61**	0.57*	0.50**	0.72**
	RMSE	17.53	21.66	20.15	24.49	22.01	19.55	17.51	19.14	19.47	14.94
TUG	r	0.26	0.05	0.35	0.34*	0.43*	0.20	0.48**	0.41	0.13	0.45*
	RMSE	6.19	6.18	5.48	5.48	5.50	6.57	5.55	6.01	6.79	5.87

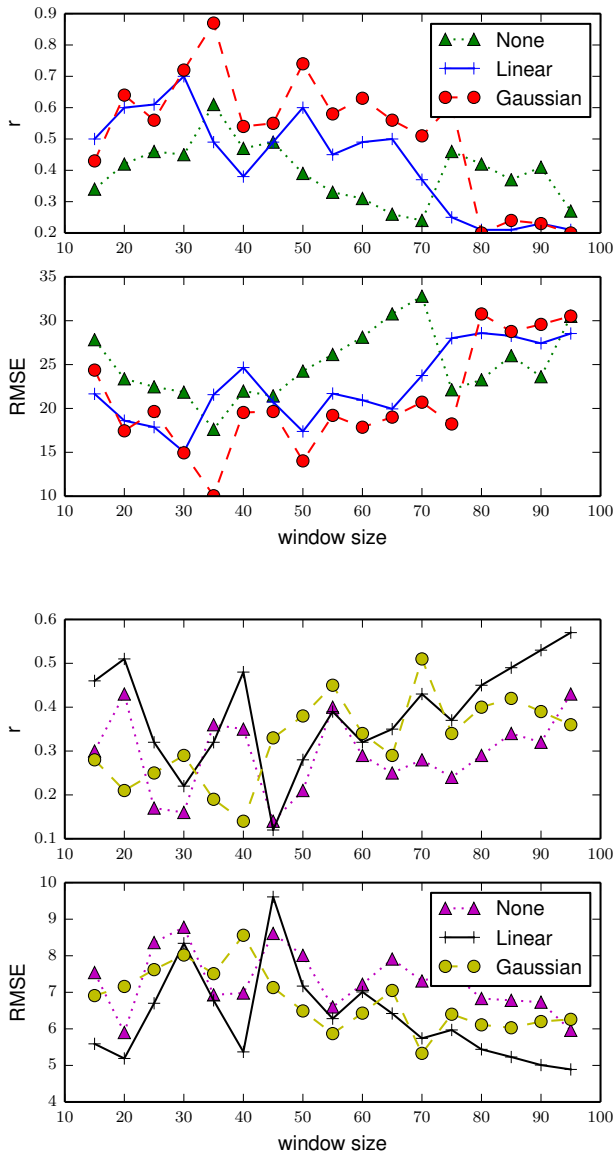


Figure 6.2: The correlation coefficients (top) and RMSE (bottom) between predicted and actual RBANS (top) and TUG (bottom) scores when we use different trend removal techniques and window sizes to train a SVR.

that the strength of the correlation coefficients is stronger and often RMSE values are smaller when we remove a Gaussian trend from the observations. Thus, in all of our remaining experiments, we remove a Gaussian trend from the data.

6.5.3 CAAB performance using random activity labels

In our final prediction experiment, we compare CAAB performance using AR-labeled activities to CAAB performance using random activity labels. There are three main objectives of this experiment. First, we want to determine the importance of the role that the AR algorithm plays in CAAB. Second, we want to verify that CAAB is not making predictions based on random chance. Third, we let prediction performance based on random activity labels serve as a baseline or lower bound performance for comparison purposes. We expect CAAB performance using AR-labeled activities to significantly outperform the baseline performance.

To perform this experiment, we create a training set in which the statistical activity features (shown in Table 6.2) are calculated from the sensor data that is randomly labeled with the activity instead of using AR algorithm to automatically generate activity labels. We performed this experiment using the following three steps: 1) We label raw sensor events by randomly choosing the activity labels from the activity set. We choose an activity assuming a uniform probability distribution

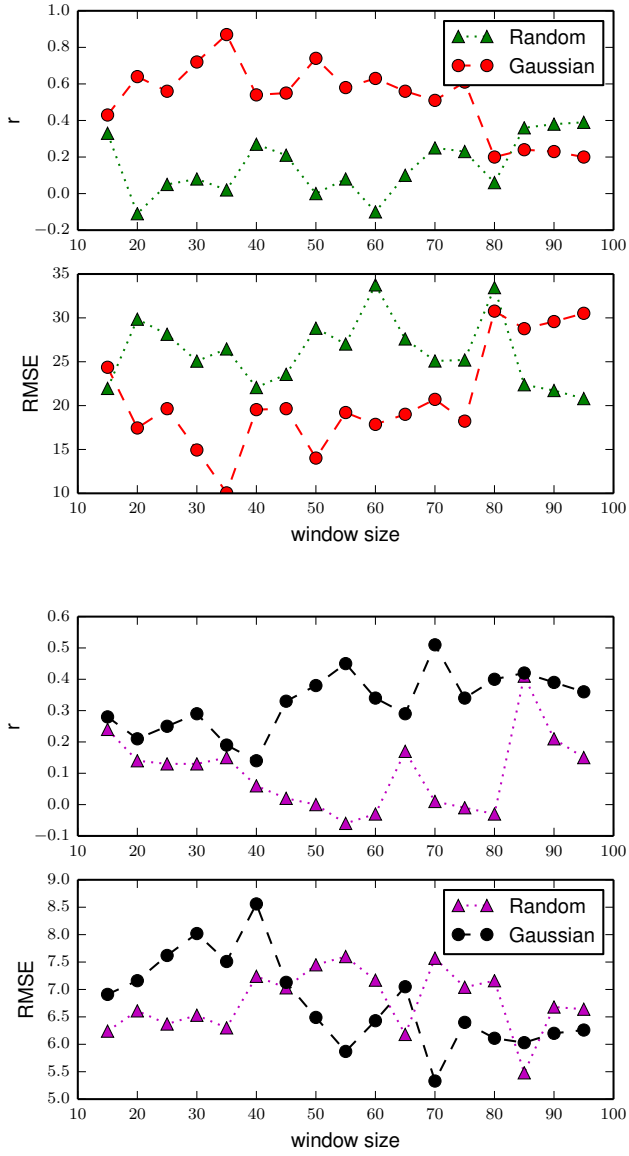


Figure 6.3: Correlation coefficients (top) and RMSE (bottom) between SVR-predicted and actual RBANS (top) and TUG (bottom) scores when we train SVR using features derived from randomly-labeled and AR-labeled activities. We use the complete set of statistical features to train the SVR.

over all activity classes. 2) We extract statistical activity features from the sensor data labeled with the random activities. 3) We train SVR using the statistical features and use clinical assessment scores as ground truth. Performance measures are computed as described in the previous sections.

As shown in Figure 6.3, we see that the strength of the correlation coefficients between predicted and actual clinical assessment scores are weak and that the RMSE values are high for the random approach. We also observed that the performances of the learning algorithms trained with features obtained from the AR labeled activities are significantly better than the random labels. Thus, we conclude that activity recognition plays a vital role in CAAB and that the CAAB predictions using statistical activity features extracted from AR labeled sensor data are meaningful and not obtained by chance.

6.6 Classification experiments

To evaluate the performance of CAAB using various classification-based experiments to evaluate, we first discretize the continuous clinical assessment scores into two binary classes and then use a learning algorithm to classify smart home residents into one of these two clinical groups. Performing these experiments allows us to use traditional supervised learning-based methods and performance measures to

evaluate CAAB, in contrast with the regression approaches that are utilized earlier in the chapter. We train the learning algorithms using the CAAB-extracted statistical activity features. For all of the classification-based experiments, we use a support vector machine (SVM) as the learning algorithm [164]. SVM identify class boundaries that maximize the size of the gap between the boundary and data points. We perform the following four different classification experiments: 1) We first evaluate classification performances of the SVM in classifying discretized RBANS and TUG clinical assessment scores when they are trained with different subsets of statistical activity features and activity performance features. 2) In the second experiment, we repeat the first experiment by discretizing RBANS and TUG scores into binary classes at different thresholds. 3) Next, we study the classification performances of the learning algorithms trained using the activity features obtained from the sensor data labeled with random activities. 4) Finally, we evaluate the classification performance (error) by using a permutation-based test to ensure that the accuracy results are not obtained by a chance.

We evaluate the classification performance of the learning algorithm using area under the curve, G-mean, accuracy and error and generate them using leave-one-out cross-fold validation.

- ROC curves assess the predictive behavior of a learning algorithm independent of error cost and class distribution. The area under the ROC curve (AUC)

provides a measure that evaluates the performance of the learning algorithm independent of error cost and class distribution.

- G-Mean is the square root of the product of the true positive and true negative rate [164].

$$G - Mean = \sqrt{(\text{true positive rate} \times \text{true negative rate})} \tag{6.1}$$

- Accuracy is the percent of the correct predictions made by the learning algorithm by the total number of predictions.

$$Accuracy = \frac{\#\text{Correct predictions}}{\#\text{Total predictions}} \tag{6.2}$$

- Error is the percent of the incorrect predictions made by the learning algorithm by the total number of predictions.

$$Error = 1 - Accuracy \tag{6.3}$$

CAAB classification performance based on feature subsets

Similar to the prediction-based experiments, we first study the importance of different subsets of statistical activity features and subsets of activities. For the first experiment, we discretize clinical assessment scores (RBANS and TUG) into binary classes using an equal frequency binning technique. We then train multiple SVMs to learn the relationship between CAAB-extracted activity features and these discretized

clinical assessment scores. We make three observations based on the classification performances presented in Tables 6.8 and 6.9.

1. From Table 6.9, we observe that the performance of the learning algorithm that is trained with the AR-labeled activities including sleep and ADLs performs generally better than using other single variables.

2. From Table 6.8, we observe that the classification performances of the SVM when trained with variance-based activity features are better for both RBANS and TUG scores. It appears that skewness-based feature is only important for classifying RBANS clinical scores and not for the TUG classifications.

3. We note that the CAAB performance in the classification-based experiments involving smart home-based parameters of sleep and ADLs are similar to the performances in the prediction-based experiments.

In the second experiment, we evaluate the impact of CAAB performance of discretizing the continuous clinical assessment scores into binary classes at different cutoff thresholds. The objective of this experiment is to identify the range of thresholds that the learning algorithm can discriminate. For this experiment, we first discretize RBANS and TUG scores into binary classes at different thresholds. We use all the features to train the SVM with AdaBoost and generate performance metrics

Table 6.8: Classification performance (accuracy and AUC) of the SVM in classifying clinical assessment scores (RBANS and TUG) discretized using equal frequency binning. We train SVM using statistical activity features from all activities.

Score Type	Measure	Change	ACF	Skewness	Kurtosis	Variance	All features
RBANS	Accuracy	26.92	57.69	73.07	57.69	63.46	71.75
	AUC	0.27	0.58	0.73	0.58	0.63	0.71
TUG	Accuracy	66.00	42.00	46.00	62.00	62.00	76.00
	AUC	0.65	0.39	0.44	0.60	0.62	0.75

using leave one out cross validation. We use SVM/AdaBoost to handle the class imbalance in the dataset if there exists one[164]. The AdaBoost algorithm improves the accuracy of the "weak" learner by assigning greater weight to the examples that the learning algorithm initially fails to correctly classify [164]. The advantages of boosting the classifier to learn an imbalanced class is that since boosting weights the samples, it implicitly performs both up-sampling and down-sampling with little information loss and is also known to prevent overfitting [164]. As showed in Figure 6.4 we observe some variations in the performance of the learning algorithms when they are trained with class labels that were discretized at different thresholds; however, the majority of the classification performances are better than random classification

Table 6.9: Classification performance (accuracy and AUC) of the SVM in classifying clinical assessment scores (RBANS and TUG) discretized using equal frequency binning. We train SVM using features from different activities.

Score Type	Measure	Sleep	Mobility	ADL	Mobility+ LeaveHome	ADL+Leave home	Sleep+ Mobility	Sleep+ ADL	Sleep+ ADL+ Leave Home	Mobility+ ADL	ALL
RBANS	Accuracy	76.92	57.69	46.15	61.53	61.53	75.00	73.08	75.00	48.05	71.15
	AUC	0.76	0.57	0.46	0.62	0.62	0.75	0.73	0.75	0.49	0.71
TUG	Accuracy	78.00	62.00	66.00	52.00	52.94	62.00	76.00	80.00	44.00	76.00
	AUC	0.77	0.61	0.64	0.52	0.50	0.62	0.75	0.79	0.43	0.75

performances (i.e., 50% accuracy for binary classes).

Additionally, based on Figure 6.4, we make four more observations:

- CAAB performance is generally better when the RBANS clinical score is discretized at thresholds within the lower range of RBANS ($85-100$) performances and within the higher range of RBANS ($125-130$) performances. It appears that the learning algorithm does successfully distinguish between the two extreme groups.

- CAAB classification performance is best when the continuous TUG clinical score is discretized at scores 12 and 17. We note that a score of 12 and above on the TUG puts individuals into the falls risk category [1]. Given that the TUG test measures the time that is required to comfortably complete the Timed Up and Go task, it appears that the learning algorithm can discriminate between the "slow performers" and the "fast performers."

- However, we note that similar to the prediction-based experiment, performance of the classifier in classifying TUG based scores is weaker than the performance while classifying RBANS scores. As we mention previously, this weaker performance is likely due to the fact that there are only two activity performance features (mobility and leave home) that represent the mobility of an individual.

- Additionally, we note that CAAB performance in classifying both TUG and

RBANS clinical labels are moderate to poor when the clinical scores are discretized into binary classes at the intermediate thresholds. We obtain moderate classification performances because the two classes are more likely to have "similar" activity performance and are therefore harder to distinguish from each other.

In the fourth experiment, we compare classification performance using AR-labeled activities and random activity labels. Similar to the prediction-based experiment, we expect the classification performance based on AR labeled activities to outperform the random method. As illustrated in Figure 6.5, we observe that AR-based classification outperforms classification with random activity labels and that the results are similar to the earlier regression-based experiments (t-test on g-mean, $p < 0.05$).

Permutation-based test

In the final experiment, we determine whether the aforementioned performance results are obtained because of chance, rather than because of the effectiveness of CAAB. With the permutation-based evaluation method, we calculate a p-value to test a null hypothesis about the relationship between the class labels and features. This p-value is calculated as a fraction of times that the performance of CAAB on the dataset that is obtained by shuffling (permuting) the class labels exceeded the performance of CAAB on the original dataset. Similar to the first classification-based

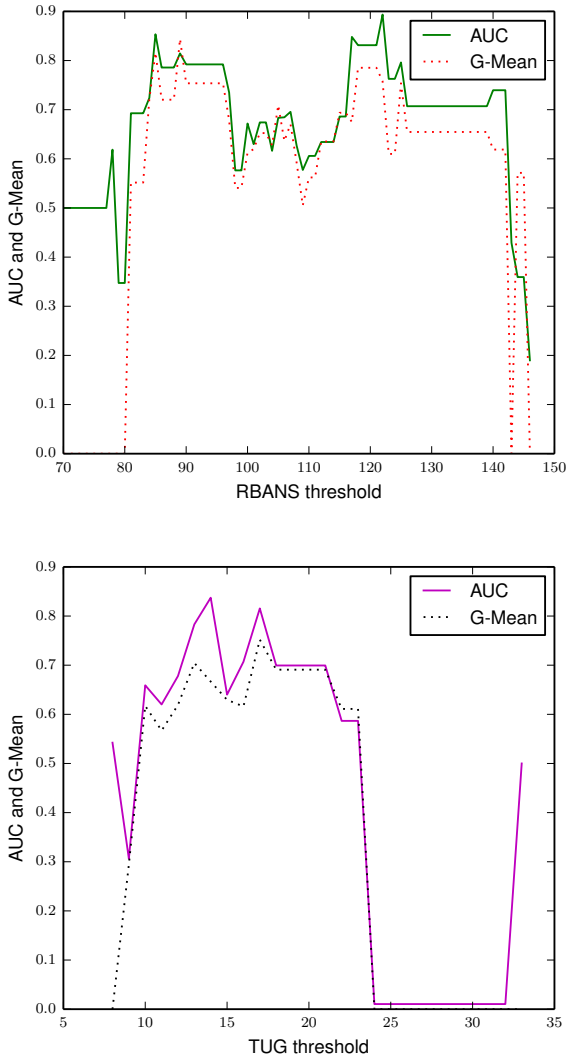


Figure 6.4: Classification performance (AUC and G-Mean) of the SVM with boosting in classifying the discretized RBANS (top) and TUG (down) scores. We discretize the RBANS score into two classes at different thresholds and train the SVM using the complete feature set.

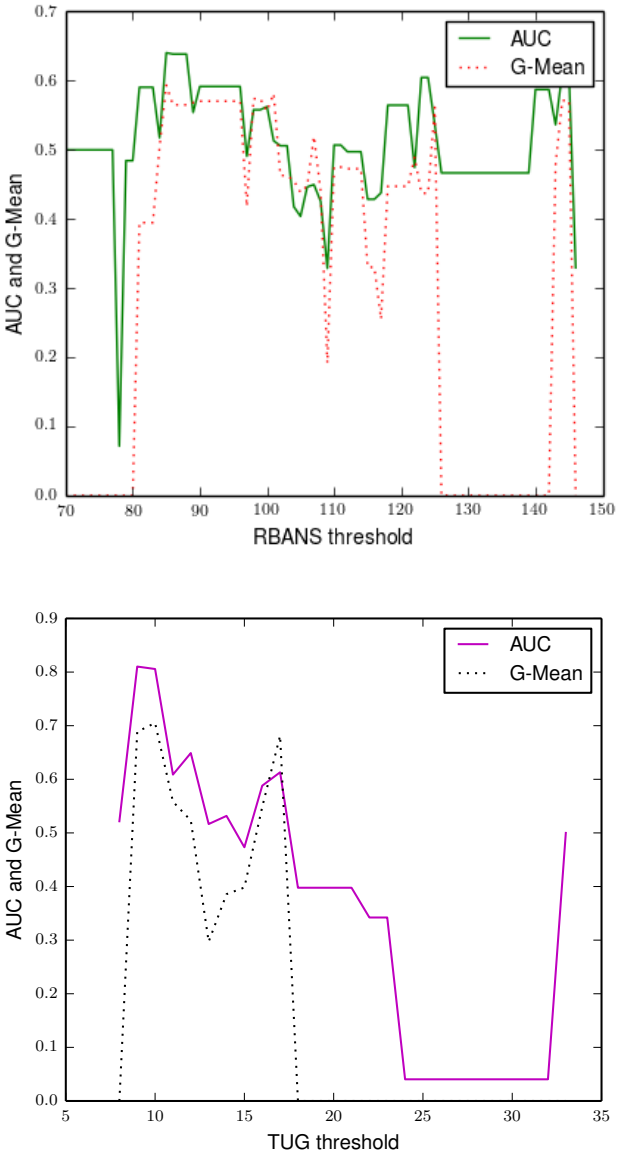


Figure 6.5: Classification performance (AUC and G-Mean) of the SVM while classifying RBANS (top) and TUG (down) clinical scores when the SVM is trained using features that are derived from randomly-annotated activities. We use the complete feature set to train the SVMs and discretize the clinical assessment scores into two classes.

Table 6.10: Average error and p-value for our test using support vector machines and activity features extracted from the dataset that is derived from AR-annotated activities (*$p < 0.05$, **$p < 0.005$)

	Original	Test 1	
Class label	Error	Err (std)	p
RBANS	0.27	0.52 (0.11)	0.009**
TUG	0.24	0.42 (0.05)	0.019*

experiment, we first discretize RBANS at a threshold of 105.5 and TUG at a threshold of 12.5 using an equal frequency binning technique. We perform a test proposed in Ojala and Garriga [104].

H: We randomly permute the class labels to study the relationship between class labels and the features. The null hypothesis is that there exists no relationship between the data and the class labels.

Table 6.10 presents the results from the AR annotated data. Based on the null hypotheses H, we make the following observation: the statistically significant ($p < 0.05$) result for the null hypothesis (H) indicates that there exists a relationship between the sensor-based activity performance and discretized RBANS and TUG labels.

Table 6.11: Average error and p-value for our test using support vector machines and activity features extracted from the dataset that is derived from randomly-labeled activities.

	Original	Test	
Class label	Error	Err (std)	p
RBANS	0.57	0.53 (0.07)	0.65
TUG	0.38	0.37 (0.11)	0.48

We repeat this experiment using activity features derived from randomly-labeled activities. Table 6.11 lists the results. Based on the p-values, we fail to reject the null hypothesis (H) that there exists no relationship between the class labels and features. Thus, we conclude that there exists a relationship between the smart home sensors-based activity features and standard clinical assessment scores (RBANS and TUG) and that the performance results are not obtained by chance.

6.7 Discussion and observations

In this chapter, we described our CAAB approach to modeling a person's activity behavior based on smart home sensor data. CAAB collects sensor data, models

activity performance, extracts relevant statistical features, and utilizes supervised machine learning to predict standard clinical assessment scores. This represents a longitudinal approach in which a person's own routine behavior and changes in behavior are used to evaluate their functional and mobility-based health. We validate our approach by performing several classification and prediction-based experiments. We found statistically significant correlations between CAAB-predicted and clinician-provided RBANS and TUG scores.

Our experiments are conducted using smart home data from 18 smart home residents and the majority of residents are cognitively healthy. Future work will include validation on larger population sizes encompassing a greater period of time. We note that CAAB is not intended to replace existing clinical measurements with the smart home-based predictions but may provide a tool for clinicians to use. We also note that an advantage of CAAB is that sparsely-measured clinical scores can be enhanced using the continuously-collected smart home data and predictions.

In the next chapter, we propose an activity curve algorithm to model behavioral daily routine of a smart home resident using longitudinal smart home sensor data. We use this activity curve model to detect changes in the daily behavioral routine and use this information to analyze changes in the cognitive and physical health of a smart home resident.

CHAPTER 7. LONGITUDINAL ANALYSIS USING ACTIVITY CURVE

The main goal of this chapter is to propose an activity curve model to represent an abstraction of an individuals normal daily behavioral routine based on automatically-recognized activities. Another goal of this chapter is to develop an algorithm to detect changes in daily behavioral routines and use this information to analyze the possibility of changes in cognitive or physical health. We will utilize longitudinal smart home sensor data collected from 18 smart home apartments to test our activity curve algorithm.

7.1 Background

Many pervasive computing applications such as home automation, activity aware interventions, and health assessment require analyzing and understanding activity-based behavioral patterns. The performance of such applications depends on the ability to correctly learn a model of general daily activity behavior from a large amount of data and be able to predict when such daily behavior is likely to continue or change. These big data-based approaches to activity modeling can then in turn be used to

provide effective activity-aware services such as improved health care.

Activity recognition lies at the heart of any pervasive computing approach to modeling behavioral routines. An activity recognition algorithm maps a sensor reading or sequence of readings to a corresponding activity label. In order to answer general questions related to daily activity patterns, such information needs to be transformed to a higher-level representation. For example, questions such as how average daily activity patterns have changed over a year, or generally what hours did a particular individual sleep last month are difficult to answer using raw output from activity recognition algorithms. However, many pervasive computing applications such as home automation and health assessment require answering such questions.

Obtaining higher-level representations or models of activities has several additional advantages. Higher-level representations can abstract variations in day-to-day activity routines. For example, wake-up times in the morning may be slightly different each day even if the overall routine is fairly stable. Additionally, such representations simplify the task of modeling an individual's daily routine and at the same time make visualization and interpretation of daily activity routines easy. Collecting big data sets over long periods of time allows us to abstract activity models over such daily variations. As we will demonstrate in this chapter, such representations aid with the process of identifying long-term changes in a behavioral routine.

For example, consider the following description highlighting aspects of an indi-

vidual's routine at two different points in time:

- *Month of March 2012*: Sleep at 10:00 PM, get up at 6:00 AM, eat breakfast at 7:00 AM, eat lunch at 12:00 PM, go out for a walk at 4:00 PM, and dine at 8:00 PM.

- *Month of September 2013*: Sleep at 8:00 PM, wake up frequently during the night, get up at 10:00 AM, no breakfast, eat lunch at 11:00 AM, no going out for a walk, and dine at 7:00 PM.

Note that each of these sample activity-based descriptions is aggregated over a one-month period and therefore describes a general routine that is maintained over a prolonged period of time. Based on these descriptions we also note changes in the routine from the first observation to the second. From this example, we can infer that by September 2013 the observed individual was experiencing disturbances in sleep, was skipping meals, and stopped exercising. Determining if the overall daily activity patterns has changed may be difficult based only on the raw sensor data or even based on event-by-event labels from an activity recognition algorithm. Such questions can be more easily answered by comparing two higher-level representations of these activity patterns.

In this chapter, we propose a novel ***activity curve*** to model an individual's generalized *daily activity routines*. The activity curve modeling algorithm uses activity-

labeled sensor events to learn a higher-level representation of the individual's regular routine. These activity labels are *automatically-recognized* using an activity recognition algorithm. We also introduce a Permutation-based Change detection in Activity Routine (PCAR) algorithm to compare activity curves between different time points in order to detect changes in an activity routine. To validate our algorithm, we make use of longitudinal smart home sensor data collected by monitoring everyday behavior of residents over two years. Finally, we demonstrate how the activity curve and the PCAR algorithm can be used to perform important pervasive computing tasks such as automated assessment of an individual's functional health.

7.2 Activity curve

An *activity curve* is a model that represents an individual's generalized activity routine. We are interested in modeling activity routines for a day-long period but this time period can be changed as needed. The activity curve uses automatically-recognized activity labels to express daily behavioral characteristics based on the timing of recognized activities.

We assume that a continuous sequence of time-stamped sensor events is available. We use an activity recognition algorithm to annotate each of these sensor events with an activity label. Activity recognition algorithms map a sequence of sen-

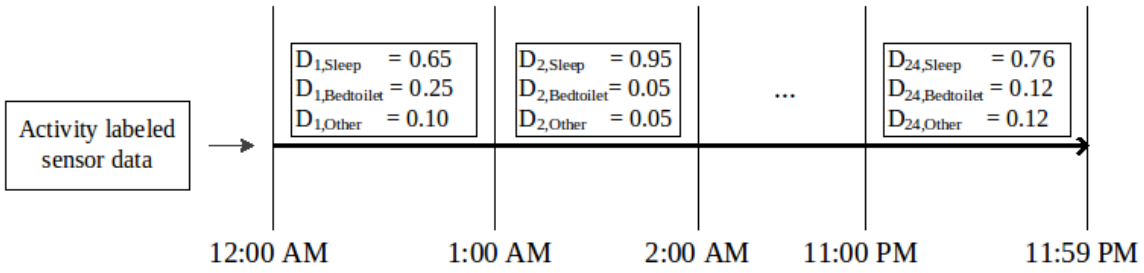


Figure 7.1: An example of activity distributions calculated at 60 minute time intervals. The figure models three possible activities: sleep, bed toilet transition, and an "other" activity. An activity curve is thus the compilation of all of these activity distributions.

sor events $\{e_1, e_2, \ldots, e_n\}$ onto the corresponding activity label A_i, where the label is drawn from the predefined set of activity classes $A = \{A_1, A_2, \ldots, A_n\}$.

We note that prevalence of common activities differs by the time of day. For example, the sleep activity dominates the prevalent distribution of activities at midnight and the cook breakfast and eat breakfast activities dominate the early morning hours. To capture such differences in activity patterns throughout the day, we segment the day-long observation period into m equal-size consecutive windows, or time intervals, and define probability distributions over activities, or activity distributions, for each of these time intervals (see Figure 7.1 for an example). An activity curve is a compilation of these activity distributions for the entire day-long period.

We also note that our activity routines tend to vary from one day to the next.

For example, we may wake up at 6:30 AM and eat breakfast at 7:15 AM one day while we might wake up at 7:30 AM and eat breakfast at 8:00 AM the next day. In order to generalize our model over such day-to-day variations in activity routines, we will define the notion of an *aggregated activity curve* that is calculated over an aggregation window of x days.

Definition 1. *Given a time interval t, an activity distribution models the daily routine based on the predefined set of activities A as a probability distribution over activities in A. The probability distribution can be estimated from sample data based on the normalized time an individual spends on a predefined set of n activities during time intervals t as observed during one or more days.*

An activity distribution for time interval t is a n-element set $\mathbf{D}_t = \{d_{t,1}, d_{t,2}, \ldots, d_{t,n}\}$ whose length is equal to that of the activity set A. The i^{th} element in an activity distribution, $d_{t,i}$, represents the probability of performing activity A_i during time interval t.

To model a person's overall daily activity routine, we use m activity distributions corresponding to each of the m time intervals. We can then construct an activity curve by collecting activity distributions that model daily activity patterns at all different times of the day.

Definition 2. *An activity curve C is the compilation of activity distributions $\boldsymbol{D}_t$ ordered by time interval t.*

The length of an activity curve is m. We refer to the model that compiles activity distributions as an "activity curve" because if we consider the activity distribution of activity A_i for all time intervals $1, 2, \ldots, m$, these activity distributions form a curve that represents the "fraction of a time" that an individual is likely to perform activity A_i over successive time intervals.

We calculate an aggregated activity distribution $\hat{\mathbf{D}}_t$ for time interval t by aggregating activity distributions $\mathbf{D}_{k,t}(1 \leq k \leq x)$ over an aggregation window of x days. If $\mathbf{D}_{1,t}, \mathbf{D}_{2,t}, \ldots, \mathbf{D}_{x,t}$ are activity distributions for the t^{th} time interval aggregated over a window of x days and they each follow a normal distribution, then we can define an aggregated activity distribution as follows.

Definition 3. *An aggregated activity distribution $\hat{\boldsymbol{D}}_t$ at time interval t is the maximum likelihood estimate of the mean that is obtained from activity distributions $\boldsymbol{D}_{k,t}(1 \leq k \leq x)$ that fall within an aggregation window of size x.*

We can write the aggregated activity distribution $\hat{\mathbf{D}}_t$ at time interval t as show in Equation 7.1.

$$\hat{\mathbf{D}}_t = \sum_{k=1}^{x} \frac{\mathbf{D}_{k,t}}{x} \tag{7.1}$$

Definition 4. *An aggregated activity curve is the compilation of aggregated activity distributions obtained over an aggregation window of size x.*

If $\Sigma = \{C_1, C_2, C_3, \ldots, C_x\}$ is a set of activity curves over an aggregation win-

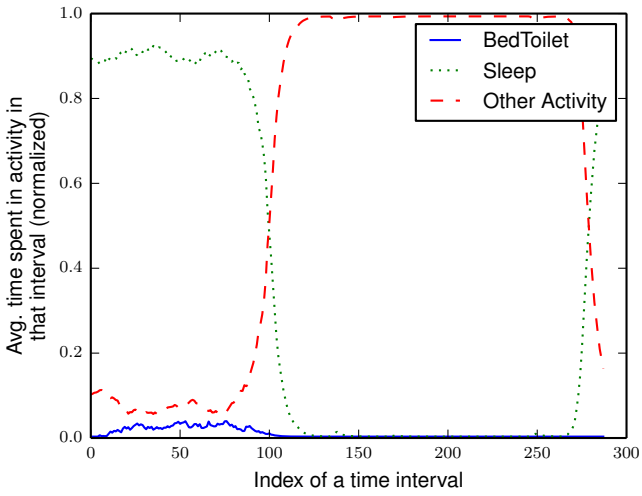


Figure 7.2: An example aggregated activity curve that models three different activities: sleep, bed toilet transition, and an "other" activity. This sample aggregated activity curve was derived using x = three months of actual smart home data. Aggregated activity distributions were calculated at 5 minute time intervals, ($m = 288$). In this graph, the time interval at index 0 represents 12:00 AM.

dow of size = x days, we can represent an aggregated activity curve over Σ as C^{Σ}. The aggregated activity curve C^{Σ} compiles the aggregated activity distributions, $\hat{\mathbf{D}}_t$. Figure 7.2 illustrates an example of an aggregated activity curve that models three different activities: sleep, bed toilet transition, and other.

7.3 Activity distribution distance

We calculate the distance between two activity distributions using the Kullback-Leibler (KL) divergence measure. We employ the KL divergence measure because it is a one of the most widely used measures to calculate the distance between two probability distributions. However, other distance metrics can also be applied for this step. Some examples of alternative measures are euclidean distance, Aitchison's distance [4], and Maximum Mean Discrepancy measure (MMD) [58]. We test one alternative distance measure, Maximum Mean Discrepancy measure, for comparison and report on the results with this measure in Section 7.8.5.

We assume that the activity distributions model the same activity set A for the same time interval size and aggregation window size. The KL divergence between two activity distributions $\mathbf{D}_1 = \{d_{1,1}, d_{1,2}, \ldots, d_{1,i}, \ldots, d_{1,n}\}$ and $\mathbf{D}_2 = \{d_{2,1}, d_{2,2}, \ldots, d_{2,i}, \ldots, d_{2,n}\}$ is defined as shown in Equation 7.2.

$$D_{KL}(\mathbf{D}_1||\mathbf{D}_2) = \sum_{i=1}^{n} d_{1,i} \log \frac{d_{1,i}}{d_{2,i}} \tag{7.2}$$

We note that the standard KL distance metric is a non-symmetric measure of the differences between two probability distributions $\mathbf{D}_1$ and $\mathbf{D}_2$. Therefore, we use a symmetric version of the Kullback-Leibler divergence between activity distributions $\mathbf{D}_1$ and $\mathbf{D}_2$, which is defined as shown in Equation 7.3. Throughout the remainder of the chapter, our discussion of KL divergence will refer to this symmetric version of

the KL divergence measure.

$$SD_{KL}(\mathbf{D}_1||\mathbf{D}_2) = D_{KL}(\mathbf{D}_1||\mathbf{D}_2) + D_{KL}(\mathbf{D}_2||\mathbf{D}_1) \tag{7.3}$$

Before defining the distance between two activity curves C_1 and C_2 of length m, we first need to align the activity distributions in the activity curves (this is described in Section 7.5). As a result of the alignment step, we obtain a vector of alignment pairs $\mathbf{\Gamma} = (p, q)$ of length $l = |\mathbf{\Gamma}|$ that aligns an activity distribution at time interval $p \quad (1 \leq p \leq m)$ of activity curve C_1 with activity distribution at time interval $q \quad (1 \leq q \leq m)$ of activity curve C_2.

We calculate the total distance, $SD_{KL}(C_1||C_2)$, between two activity curves, C_1 and C_2, as the sum of distances between each aligned activity distribution for the two activity curves, as shown in Equation 7.4.

$$SD_{KL}(C_1||C_2) = \sum_{\alpha=1}^{l} SD_{KL}(\mathbf{D}_{1,p}||\mathbf{D}_{2,q}) \text{ such that } \mathbf{\Gamma}_\alpha = (p, q) \tag{7.4}$$

where $\mathbf{D}_{1,p}$ and $\mathbf{D}_{2,q}$ are the activity distributions that belong to activity curves C_1 and C_2 at time intervals p and q, respectively.

7.4 Determining the size of an aggregation window

Daily activity routines are performed differently from one day to the next. As a result, the daily activity curve that models these activity routines will vary from one

day to the next. We want to calculate an aggregated activity curve that generalizes over minor day-to-day variations while still capturing the typical routine behavior. When determining the appropriate size of an aggregation window, our goal is to find the smallest possible number of days that is considered stable. We determine that an aggregated activity curve is stable if the shape of the curve remains mostly unchanged when more days are added to the aggregation window. By keeping the aggregation window small, our model can be more sensitive to significant changes in routine behavior. If the window is too small it will not be general enough to encompass normal variations in daily routines. We propose Algorithm 4 to determine the minimum length of an aggregate window that is required to calculate a stable, representative aggregated activity curve for a particular time interval. We choose the minimum aggregate window size x_{min} such that no smaller window would ensure the stability criterion.

To determine the ideal aggregation window size, we start with a window of size $x = 2$ and consider the corresponding aggregated activity curve C^{Σ}, aggregated from the set of individual activity curves $\Sigma = \{C_1, \ldots, C_x\}$. We estimate the distance between C^{Σ}_x and C^{Σ}_{x+1}. If the distance is greater than a predefined threshold T, we increase the window size. Therefore, if $SD_{KL}(C^{\Sigma}_x || C^{\Sigma}_{x+1}) < T$ and $SD_{KL}(C^{\Sigma}_{x+1} || C^{\Sigma}_{x+2}) < T$, then x is selected as the representative aggregation window size, otherwise we increase size of the aggregation window by one and repeat the

process. This process is shown in Algorithm 4.

Algorithm 4 AggregationSize(Σ)

1: // Calculate the minimum size of an aggregation window.

2: $\Sigma = \{C_1, C_2, \ldots, C_N\}$ for each of the N days in the input data.

3: // Return the minimum aggregation window size.

4: **initialize** $x = 2$

5: **repeat**:

6: Create C_x^{Σ}, aggregated activity curve for window size x.

7: Create C_{x+1}^{Σ}, aggregated activity curve for window size $x + 1$.

8: Create C_{x+2}^{Σ}, aggregated activity curve for window size $x + 2$.

9: Compute d_1 = distance between two aggregated activity curves $SD_{KL}(C_x^{\Sigma}||C_{x+1}^{\Sigma})$.

10: Compute d_2 = distance between two aggregated activity curves $SD_{KL}(C_{x+1}^{\Sigma}||C_{x+2}^{\Sigma})$.

11: **if** $d_1 < T$ and $d_2 < T$ $x = x + 1$

12: **else** return x

13: **until** $x < N$

7.5 Activity curve alignment

In order to compute similarity (or distance) between two activity curves, we need to compare each of the activity distributions that belong to these two activity curves. However, we first need to determine which pairs of distributions to compare by considering alternative distribution alignment techniques. Activity curve alignment can be performed based on aligning the same time of day between two curves. Alternatively, we can try to maximally align the activity occurrences between two curves before performing such a comparison. Here we provide details for these two alignment techniques that we use in our work.

7.5.1 Time interval-based activity curve alignment

The time interval-based activity curve alignment technique presumes that distributions between two curves should be aligned based on time of day and thus aligns activity distributions between two activity curves if the time intervals are the same. In essence, this method does not make any extra effort to align activities that occur at different times in the distribution, but simply compares the activity distributions based on time of day alone.

If C_1 and C_2 are two activity curves of length m, the time interval-based activity distribution alignment method aligns the corresponding activity distributions using

a vector of alignment pairs, $\mathbf{\Gamma} = (r, s)$, such that $r = s$. This technique aligns an activity distribution at time interval $r \quad (1 \leq r \leq m)$ of activity curve C_1 with activity distribution at time interval $s \quad (1 \leq s \leq m)$ of activity curve C_2.

7.5.2 Dynamic time warping-based activity curve alignment

A person's routine may be relatively stable, even though there are minute changes in the time an activity occurs or the duration of a particular activity. For example, an individual may sleep at 10:00 PM one day, an hour earlier at 9:00 PM the next day, an hour later at 11:00 PM a few days later, and eventually go back to sleeping at 10:00 PM. Aligning activity distributions using dynamic time warping allows us to maximally align common activities before comparing two activity curves. Such an alignment accommodates activity time changes that are shifted temporally backward (for example, an hour earlier), forward (for example, an hour later), expanded (longer duration), compressed (shorter duration), or not changed at all from one day to another. We optimize activity alignment using Dynamic Time Warping (DTW) to align distributions between two activity curves.

Dynamic time warping finds an optimal alignment or warping path between activity curves. This optimal warping path has minimal total cost among all possible warping paths. We use the symmetric KL distance metric that we previously men-

tioned to compute this warping path. The warping path has the following three main properties:

- *Boundary property*: The first and last elements (activity distributions) from the two activity curves are always aligned with each other.
- *Monotonicity property*: Paths are not allowed to move backwards.
- *Step size property*: No activity distributions are omitted from the curve alignment.

We also note that due to the monotonicity property, DTW does not allow backward alignments. However, as we have seen in practice, activity distributions can be shifted temporally backward and/or temporally forward. Therefore, we modify the standard approach to perform two independent iterations of DTW:

- In *forward dynamic time warping*, we start from the first activity distribution and move forward in time toward the last activity distribution to find an optimal alignment between activity curves that are similar in the forward time direction.
- In *backward dynamic time warping*, we start from the last activity distribution and move backward in time toward the first activity distribution to find an optimal alignment between activity curves that are similar in the backward time direction.

If C_1 and C_2 are two activity curves of length m, the DTW-based activity distribution alignment outputs two alignment vectors, $\mathbf{\Gamma}_{forward} = (u, v)$ of length $l_{forward}$, and $\mathbf{\Gamma}_{backward} = (r, s)$ of length $l_{backward}$, respectively. The forward DTW aligns an activity distribution from curve C_1 at time interval $u \quad (1 \leq u \leq m)$ with an activity distribution from curve C_2 at time interval $v \quad (1 \leq v \leq m)$. Similarly, the backward DTW aligns an activity distribution from curve C_1 at time interval $r \quad (1 \leq r \leq m)$ with an activity distribution from curve C_2 at time interval $s \quad (1 \leq s \leq m)$. The DTW method outputs whichever vector, $\mathbf{\Gamma}_{forward}$ or $\mathbf{\Gamma}_{backward}$, that results in the maximal alignment between the two distributions and thus minimize the difference. We will utilize these two different alignment techniques in our PCAR algorithm to detect changes between two aggregated activity curves and calculate change scores.

7.6 PCAR

Based on our notion of an activity curve, we now introduce our Permutation-based Change Detection in Activity Routine (PCAR) algorithm. This algorithm identifies and quantifies changes in an activity routine. PCAR operates on the assumption that daily activities are scheduled according to a routine and are not scheduled randomly. For example, we regularly "wake up","bathe" and "have breakfast" in the morning and "dine" and "relax" in the evening. In contrast, we rarely dine in the

middle of night. Such regularities are useful, for example, to determine if there are significant changes in lifestyle behavior that might indicate changes in cognitive or physical health.

7.6.1 Permutation-based two-sample test

PCAR identifies significant changes in an activity routine using a two-sample permutation test [52]. The permutation-based technique provides a data-driven approach to calculate an empirical distribution of a test statistic. The empirical distribution of a test statistic is obtained by calculating the test statistic after randomly shuffling (rearranging) the data a specified number of times. The permutation-based test is *exact* if the joint distributions of rearranged samples are the same as the joint distribution of the original samples. In other words, the samples are *exchangeable* when the null hypothesis is true. This type of test allows us to determine the significance of a difference between two aggregated activity curves.

We use a permutation-based test to perform a two-sample homogeneity test. In a two-sample homogeneity test, we test the null hypothesis that the two samples come from the same probability distribution versus the alternate hypothesis that they come from different probability distributions. There are three main steps involved in the permutation-based two-sample test.

- *Calculate the test statistic*: Compute the test statistic from the original samples.
- *Permutation*: Rearrange the samples and compute the test statistic again. Repeat this step a specified number of times to obtain an empirical distribution of the test statistic.
- *Significance testing*: Compare the test statistic obtained from the original (unpermuted) set of data with the empirical distribution of the test statistic from the permuted data to calculate the *p-value*. The *p-value* is calculated based on the relative ranking of the original test statistic in the empirical distribution of the test statistic (i.e., the ratio of the number of times the test statistic from the permuted sample is equal to or greater than the original test statistic to the total number of permutations).

If the *p-value* is significant (i.e., $\alpha < 0.01$), the null hypothesis is rejected in the favor of the alternative hypothesis.

7.6.2 Changes in activity distributions

We use the permutation-based two-sample test to determine whether there is a significant change among a set of activity distributions at a particular time interval. We formulate the null hypothesis that the set of activity distributions comprising two activity curves are identical versus the alternative hypothesis that the set of activity

distributions is significantly different between the two aggregated activity curves.

We test the hypothesis of a significant change between two aggregated activity distributions, $\hat{\mathbf{D}}_{1,t}$ and $\hat{\mathbf{D}}_{2,t}$.

- *Calculate the test statistic*: Calculate the test statistic $Dist_t = SD_{KL}(\hat{\mathbf{D}}_{1,t}||\hat{\mathbf{D}}_{2,t})$ between the two aggregated activity distributions.

- *Permutation*: Rearrange the order of individual activity distributions that comprise each of the aggregated distributions and recalculate the aggregated distributions $\hat{\mathbf{D}}_{1,t}$ and $\hat{\mathbf{D}}_{2,t}$. Calculate the KL divergence between the new aggregated activity distributions $\hat{Dist}_t = SD_{KL}(\hat{\mathbf{D}}_{1,t}||\hat{\mathbf{D}}_{2,t})$. Repeat the process a specified number of times to obtain an empirical distribution of KL divergence (the test statistic), $\hat{Dist}_t$.

- *Significance testing*: To test if a significant difference exists between $\hat{\mathbf{D}}_{1,t}$ and $\hat{\mathbf{D}}_{2,t}$ calculate the *p-value* based on the ranking of the original test statistic $Dist_t$, in the empirical distribution $\hat{Dist}_t$. If a small *p-value* is obtained, reject the null hypothesis in favor of alternative hypothesis. This is shown in Equation 7.5.

$$p_{perm} = \frac{\#\hat{Dist}_t > Dist_t}{\#\text{permutations}} \tag{7.5}$$

where $\hat{Dist}_t$ is the empirical distribution of the test statistic at the t^{th} time interval. We reject the null hypothesis that no changes have occurred at a significance level of

$\alpha = 0.01$.

7.6.3 Changes in activity curves

We now extend the technique of detecting significant changes between activity distributions to quantify the difference in activity routine observed from two separate aggregation windows, W_1 and W_2, each of size x days. To do this, PCAR counts the total number of significant differences between the individual activity distributions within window W_1 and the distributions within window W_2 to output a *change score* that quantifies the significant changes observed among the activity curves.

PCAR calculates a sum, S, over changes that are detected between activity curve distributions for each individual time interval. In order to identify the time intervals at which changes in the activity distributions comprising the aggregated activity curves are deemed significant, PCAR performs the following steps:

- *Permutation*: Calculate the empirical distributions of the test statistic (KL divergence) by permuting and comparing the individual activity distributions within the two aggregation windows W_1 and W_2 at each time interval using the method summarized in Algorithm 5.

- *Alignment*: Calculate the two aggregated activity curves C_1 and C_2 using the activity distributions aggregated for each time interval over windows W_1 and

W_2. Align curves C_1 and C_2 using one of the alignment techniques described in the previous section to generate the alignment vector $\mathbf{\Gamma} = (u, v)$.

- *Calculate the test statistic*: For each alignment pair $(u, v) \in \mathbf{\Gamma}$, calculate the test statistic $Dist_{u,v}$ between the aggregated activity distribution at time interval u of activity curve C_1 and the aggregated activity distribution at time interval v of C_2.

- *Significance testing*: To test if there is a significant change between activity distributions at time intervals u and v, calculate the *p-value* based on the relative ranking of $Dist_{u,v}$ in the empirical distributions. The steps are summarized in Algorithm 6.

We note that we compare at least m activity distributions during this process where m is the number of aggregated activity distributions in an activity curve. To control the False Discovery Rate (FDR) at level $\alpha^*(\alpha^* = 0.01)$, we apply the Benjamini-Hochberg (BH) method [15]. The BH method first orders the *p-values*, $p_{(1)}, p_{(2)}, \ldots, p_{(k)}, \ldots, p_{(m)}$, in ascending order and for a given value of α^*, the BH method finds the largest k such that $p_{(k)} \leq k \times \frac{\alpha^*}{m}$. The BH algorithm rejects the null hypothesis corresponding to $p_{(i)}$ if $i \leq k$. If a significant change is detected between aligned activity distributions, PCAR increments its change score, S, by one. PCAR generates two different change scores based on the alignment techniques that

are employed: either the same index alignment or the DTW-based alignment.

7.7 Use of activity curves for smart functional assessment: A case study

An activity curve model provides a big data-based tool for representing a longer-term behavioral model. Such a tool is valuable for a variety of applications including human automation, health monitoring, and automated health assessment. In this section, we explain how the activity curve model and the PCAR algorithm can be instrumental in performing automated functional assessment.

Activities of daily living such as sleeping, grooming, and eating are essential everyday functions that are required to maintain independence and quality of life. Decline in the ability to independently perform these ADLs has been associated with a host of negative outcomes, including placement in long-term care facilities, shortened time to conversion to dementia, and poor quality of life for both the functionally impaired individuals and their caregivers [86, 89, 105].

We use smart home sensor data to derive activity curves that model the activity routines of a smart home resident. Our PCAR algorithm detects changes in those activity routines. We then analyze the relationship between standard clinical scores and detected changes in ADL patterns. To validate our automated assessment technique, we utilize smart home sensor data that was collected from real world smart home

Algorithm 5 EmpiricalDistribution(Σ_1, Σ_2, N_P)

1: // Build empirical distribution $\hat{Dist}$ of the test statistic.

2: Σ_1, Σ_2 = two sets of activity curves

3: N_p = number of permutations

4: **initialize** $\hat{Dist}$ as $N_p \times m$ matrix $\triangleright$ m is # activity distributions in the activity curves

5: **initialize** $i = 0$

6: **while** $i < N_p$ **do** :

7: Rearrange the activity curves.

8: Generate aggregated activity curves C^{Σ_1} and C^{Σ_2} by aggregating the distributions in Σ_1, Σ_2

9: Using the time interval-based alignment technique, align the two aggregated activity curves to obtain an alignment vector $\mathbf{\Gamma}$.

10: **for all** alignment pairs (u, u) in $\mathbf{\Gamma}$ **do** :

11: Find a distance $SD_{KL}(\mathbf{D}_{1,u}||\mathbf{D}_{2,u})$ between u^{th} activity distributions in two activity curves.

12: Insert $SD_{KL}(\mathbf{D}_{1,u}||\mathbf{D}_{2,u})$ to empirical distribution $\hat{Dist}$ at location $[i, u]$.

13: **end for**

14: i = i+1

15: **end while**

16: return $\hat{Dist}$

testbeds with older adult residents. We apply robust activity recognition algorithms [33, 78] to label these sensor-monitored data with the corresponding activity labels.

Algorithm 6 PCAR($\Sigma_1, \Sigma_2, \hat{Dist}$)

Σ_1, Σ_2 = two sets of activity curves

//Return a change score S

C_1 = AggregateActivityCurves(Σ_1)

C_2 = AggregateActivityCurves(Σ_2)

$\mathbf{\Gamma}$ = AlignCurves(C_1, C_2)

for all alignment pairs (u, v) in $\mathbf{\Gamma}$ **do** :

    Calculate $SD_{KL}(\mathbf{D}_{1,u}||\mathbf{D}_{2,v})$ between activity distribution $\mathbf{D}_{1,u} \in C_1$ and $\mathbf{D}_{2,v} \in C_2$.

    Perform significance testing of estimated distance by querying $\hat{Dist}$.

    if change is significant :

    $S = S + 1$

end for

return S

7.7.1 Synthetic dataset

First, we validate the performance of the proposed PCAR algorithm by running it on a synthetic activity curve. We create a synthetic activity curve by compiling synthetic activity distributions. The synthetic activity distribution models the patterns of two activities, an arbitrary activity A and an "other" activity.

We generate synthetic activity distributions for each time interval t for N days by applying the following three steps. Here l represents the length of each time interval.

- Generate a random value $p \quad (0 \leq p \leq l)$, which represents the average time that is spent in performing activity A during time interval t.

- Generate two vectors, S and S', each of length N. Generate the vector S from a normal distribution i.e. $S \sim \text{Normal}(p, 1)$. Each element of vector S' is generated by subtracting the corresponding value in S from l i.e $l - s \in S'$.

- Create an activity distribution that models patterns of two activities at a time interval t. The elements of the activity distribution is $[s, l - s]$. We combine these individual synthetic activity distributions at different time intervals into activity curves. When we introduce an activity change, we multiply the average time spent performing an activity A by a constant factor in each time interval.

Using the aforementioned method, we create two different sets of activity curves. The first set Y does not contain any changes and another different set of daily activity curves Z contains changes in all activity distributions every 90 days. We run the PCAR algorithm on an aggregated activity curve of size 90 days using the time interval index alignment. PCAR successfully detects all changes in the dataset where activity changes were made and does not produce any false positives in the synthetic dataset that does not contain changes.

7.7.2 Experimental setup

Next, we demonstrate how PCAR can be used to detect changes in real smart home data. For this study, we recruited 18 single-resident senior volunteers from a retirement community and installed smart home sensors in their homes [29]. The smart home sensors unobtrusively and continuously monitor resident activities. We continuously collected raw sensor events for an extended period ($\sim$ 2 years) from all of the residents. At the same time, standardized clinical, cognitive, and motor tests were administered biannually to the residents. Refer to Section 2.1.3 for more details on CASAS longitudinal smart home testbed.

7.7.3 Participants

Participants included 18 senior residents (5 females, 13 males) from a senior living community. All participants were 73 years of age or older, and had a mean level of education of 17.52 years. At baseline, participants were classified as cognitively healthy ($N = 7$), at risk for cognitive difficulties ($N = 7$) or experiencing cognitively difficulties ($N = 4$). One participant in the cognitively compromised group met the criteria for dementia [10], while the other three individuals met the criteria for mild cognitive impairment (MCI) [5].

7.7.4 Smart home testbed

The 18 smart home testbeds are single-resident apartments, each with at least one bedroom, a kitchen, a dining area, and one bathroom. For more details about these smart apartments, see Section 2.1.3.

The residents perform their normal activities in their smart apartments, unobstructed by the smart home instrumentation. While residents carry out their daily routines, sensors continuously monitor their behavior. The middleware collects the sensor events and stores them on a database server. Each sensor event is represented by four fields: date, time, sensor identifier, and sensor message. The raw sensor data does not contain activity labels. We use an activity recognition algorithm, described

in the next section, to map individual sensor events to corresponding activity labels as shown in Figure 2.7.

7.7.5 Activity recognition

Activity recognition algorithms label activities based on readings (or events) that are collected from the smart environment, mobile device, or other sensors. As described earlier, the challenge of activity recognition is to map a sequence of sensor events onto a value from a set of predefined activity labels. These activities may consist of simple ambulatory motion, such as walking and sitting, or complex activities of daily living, such as cooking and eating, depending upon what type of underlying sensor technologies and learning algorithms are used. We use our AR activity recognition algorithm, described in Section 2.1.6, to label individual sensor events with corresponding activity labels as shown in Figure 2.7.

7.7.6 Activities

Using the AR algorithm, we recognize seven different activities of daily living: sleep, bed-to-toilet, cook, eat, personal hygiene, leave home, relax, and an "other" activity. In our datasets, the relax activity includes watching television, reading, and/or napping that typically takes place in a favorite chair or location where the

resident spends time doing these activities. We note that these activities are separately considered for the longitudinal study described in Chapter 6.

One of the activities that we assess in this case study, sleep, is a particularly important clinical construct that both clinicians and caregivers are interested in understanding [43]. Sleep problems in older adults can affect cognitive abilities [69, 134] and have been associated with decreased functional status and quality of life [49, 90]. Moreover, individuals with dementia often experience significant disruption of the sleep-wake cycle [51].

On the other hand, all basic activities of daily living (e.g., eating, grooming) and instrumental activities of daily living (IADLs; e.g., cooking, managing finances) are fundamental to independent living. We postulate that like sleep, changes in any of these activities may indicate changes in cognitive or physical health. Data indicate that subtle difficulties in everyday activity completion (e.g., greater task inefficiencies, longer activity completion times) occur with increased age [131]. Clinical studies have also demonstrated that individuals diagnosed with MCI experience greater difficulties (e.g., increased omission errors) completing everyday activities when compared with healthy controls [9, 53, 110, 133]. Furthermore, with incidence of severe cognitive problems such as AD, individuals have difficulty in both initiating and completing basic activities [59]. Thus, clinicians argue the importance of understanding the course of functional change given the potential implications for developing methods

for both prevention and early intervention [130, 131].

Finally, our AR algorithms places all other activities into an "other" category. This category includes all remaining activities that we are not distinguishing and analyzing separately in this case study. Thus, in our current work, we analyze daily patterns of 7 activities of daily living and the "other" activity using activity recognition algorithms and analyze them using the PCAR algorithm.

7.7.7 Standard neuropsychological tests

Clinical tests were administered every six months to residents of our smart home testbeds. As detailed in Table 2.3, these tests included Timed Up and Go Test (TUG) and a global measure of cognitive status (RBANS). The administered clinical tests are standardized and validated measures that provide indication of mobility-based health and cognitive health. Using repeated measurements obtained from biannual clinical tests, we create a clinical longitudinal dataset that contains these two measurement variables (features).

7.8 Experimental results

7.8.1 Preprocessing

We use two different preprocessing techniques to preprocess activity curves.

- Mean smoothing: We run a mean smoothing filter of size 3 on activity distributions comprising the activity curve to smooth out noise and minor variations. In this step, we replace the estimate at time interval t with the average estimate of activity distributions at times $t-2, t-1,$ and t.

- Add-One smoothing: Activity distributions for certain activities can be zero. For example, we rarely eat and cook at midnight so activity distributions of these activities at midnight are often zero. We perform add-one smoothing on all of the elements of activity distributions. In add-one smoothing, we add a constant $\alpha = 1$ in every elements of the activity distributions. Add-one smoothing technique is often used in natural language processing to smooth unigram estimates and has an effect of removing zero entries.

7.8.2 Studying aggregated activity curve

Figure 7.3 is an example aggregated activity curve that models eight different activities, including seven recognized activities(i.e., sleep, bed toilet transition, eat,

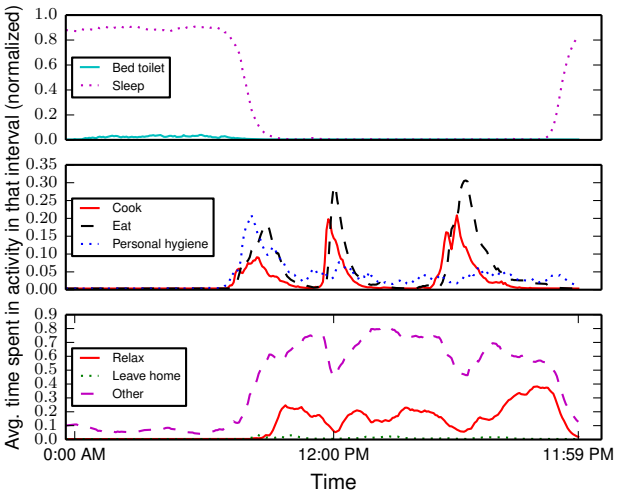


Figure 7.3: An example of aggregated activity curve that models eight different activities. This sample aggregated activity curve was derived using x = three months of actual smart home data. Aggregated activity distributions were calculated at 5 minute time intervals, ($m = 288$)

cook, relax, personal hygiene) and an "other" activity. This sample aggregated activity curve was derived using an aggregation window of size x = three months based on actual single resident smart home sensor data and 5 minute time intervals. We observe that this smart home resident usually goes to sleep at around 9:00 PM and wakes up at around 7:00 AM. We also observe that the resident exhibits a fairly fixed schedule for eating breakfast, lunch, and dinner.

A primary contribution of this research is the introduction of an activity curve as a tool for representing behavior aggregated over a specific time period. Therefore, we want to confirm that the representation is accurate and that it captures the true behavior of a smart home resident. To perform this experiment, we visually inspect an aggregated activity curve to obtain the corresponding activity start and end times. Then, we compare these activity start/end times with the activity start/end times obtained from manually annotated sensor data. To quantify the accuracy, we calculate the absolute difference between the respective start and end times obtained from these two different sources. We consider a time difference of 30 minutes as an acceptable error margin. This size error margin means that if an activity curve represents an individual's sleep start time at 10:00 PM, his true sleep start time is sometime between 9:30 P.M and 10:30 P.M. We consider such a difference as an acceptable given that the time individuals typically initiate and complete daily activities (e.g., sleep) varies to at least this extent.

To perform this experiment, we use seven days of AR-annotated sensor data and generate aggregated activity curves of two randomly chosen residents (see Figure 7.4). We visually inspect the activity curves and note the most likely start and end times of the activities. We chose a set of four activities sleep, eat, cook, and bed toilet transition. We note that individuals typically initiate and complete three of the four activities (sleep, eat, cook) almost at around a fixed time of a day. While we note start time and end time for the sleep activity, we note three different start and end times of the cooking and eating activity (cook/eat breakfast, lunch and dinner). We note the first occurrence for the bed and toilet transition activity.

Second, we followed the same procedure to obtain the activity start and end times from the manually annotated sensor data. If there are multiple annotations of the same activity, we consider the first annotation of the activity. For each activity, we find an average start and end time.

Next, for each activity, we calculate the absolute difference between activity curve-based activity start time and manually annotated sensor data-based start time. Similarly, we repeat the previous step for the activity end times. We find that the mean absolute difference is about 26 minutes. The activity start/end times obtained from activity curve representation and manually annotated sensor data with the corresponding calculations are listed in Table 7.1. Since, the absolute mean difference of 26 minutes falls below our acceptable error margin of 30 minutes, we conclude that

Table 7.1: Activity start and end times obtained from activity curve representation and manually annotated sensor data.

	Resident I			Resident II		
	Activity curve-based time	Manually annotated time	Difference (minutes)	Activity curve-based time	Manually annotated time	Difference (minutes)
Sleep start	10:15:00 PM	10:50:00 PM	35	09:30:00 PM	10:03:00 PM	33
Sleep end	08:15:00 AM	08:04:00 AM	11	06:30:00 AM	05:45:00 AM	45
Eat breakfast start	08:30:00 AM	08:54:00 AM	24	06:30:00 AM	07:02:30 AM	32
Eat breakfast end	09:00:00 AM	09:20:00 AM	20	07:00:00 AM	07:11:00 AM	11
Eat lunch start	12:30:00 PM	12:22:00 PM	8	11:30:00 AM	11:56:00 AM	26
Eat lunch end	01:00:00 PM	12:39:00 PM	21	12:30:00 PM	12:33:00 PM	3
Eat dinner start	05:30:00 PM	06:03:00 PM	33	04:30:00 PM	04:50:00 PM	20
Eat dinner end	07:00:00 PM	06:34:00 PM	26	05:30:00 PM	05:21:00 PM	9
Cook breakfast start	08:30:00 AM	08:44:00 AM	14	07:00:00 AM	06:33:00 AM	27
Cook breakfast end	09:00:00 AM	08:54:00 AM	6	07:30:00 AM	07:04:00 AM	26
Cook lunch start	12:30:00 PM	12:23:00 PM	7	11:30:00 AM	11:34:00 AM	4
Cook lunch end	01:00:00 PM	12:31:00 PM	29	12:30:00 PM	12:19:00 PM	11
Cook dinner start	05:30:00 PM	05:48:00 PM	18	04:15:00 PM	04:42:00 PM	27
Cook dinner end	06:30:00 PM	06:21:00 PM	9	05:30:00 PM	05:16:00 PM	14
Bed to toilet start	02:00:00 AM	03:09:00 AM	69	01:30:00 AM	02:48:00 AM	78
Bed toilet end	02:15:00 AM	03:15:00 AM	60	01:30:00 AM	02:49:00 AM	79
		Mean	24.37		Mean	27.81

activity curve correctly represents the activity start and end times of the activities.

7.8.3 Comparing activity distributions

For most individuals, our activities follow a common pattern based on factors such as time of day. Thus, the activity distributions that model activities at different time intervals belonging to different times of a day will be different. In our first ex-

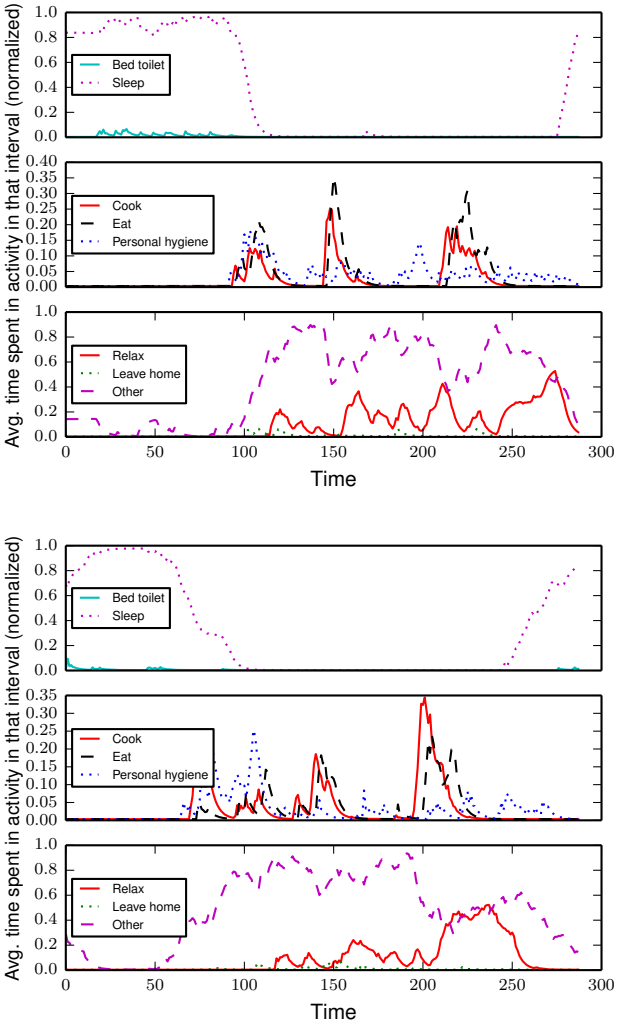


Figure 7.4: Examples of aggregated activity curves that model eight different activities. These sample aggregated activity curves were derived using seven days of AR annotated smart home data. Aggregated activity distributions were calculated at 5 minute time intervals ($m = 288$).

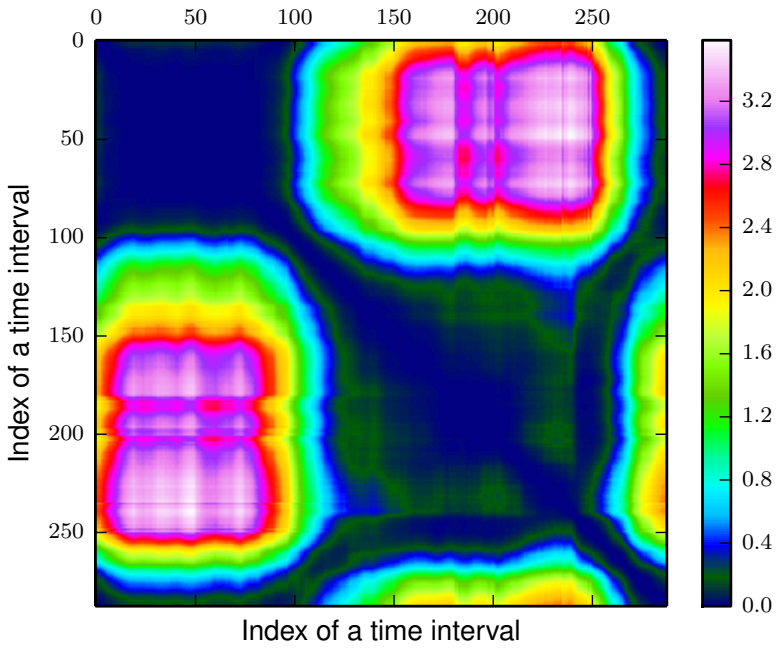


Figure 7.5: Heat map of the pairwise distance matrix between activity distributions of an aggregated activity curve using KL distance. The size of the time interval is 5 minutes. Index 0 represents 12:00 AM.

periment, we assess whether our proposed activity curve can capture such differences in activity distributions. We calculate an aggregated activity curve for five-minute time intervals using the first three months of activity-annotated sensor data from one of our smart homes. We calculate a pairwise distance (symmetric KL divergence) matrix between activity distributions from this aggregated activity curve. We plot this pairwise distance matrix in a heat map shown in Figure 7.5.

From the heat map in Figure 7.5, we observe that the distance between activity distributions varies according to the time of day. We observe that the darkest colors appear along the diagonal when we compare activity distributions for the same time of day. In contrast, we observe that the hottest colors (greatest distance) occurs when comparing activities at midnight (when the resident typically sleeps) to activities in mid-afternoon when the resident is quite active. Additionally, we also see different clusters emerge (for instance, between times 0 and 100) corresponding to times of day. These observations provide intuitive visual evidence that the activity curve is capturing generalizable differences in activity routine at various times of the day.

In the next experiment, we study how the activity distribution distances within an activity curve (the y axis in Figure 7.6) change as a function of the time interval size (the x axis). For this experiment, we calculated an average pairwise distance between activity distributions within aggregated activity curve for each time interval size. We observe that as the time interval increases, the average pairwise distance between daily activity distributions decreases. Such a decrease in distances is observed because activity distributions at larger sized time intervals are overwhelmed by activities that take larger duration (such as sleep). As a result, smaller differences between such activity distributions are harder to detect.

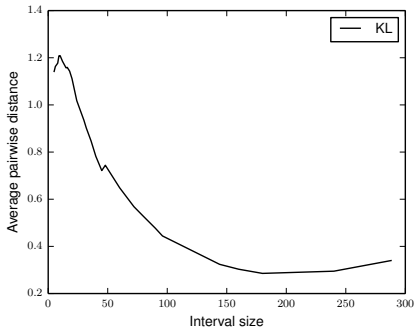


Figure 7.6: Average intra-curve pairwise KL distance as a function of time interval size.

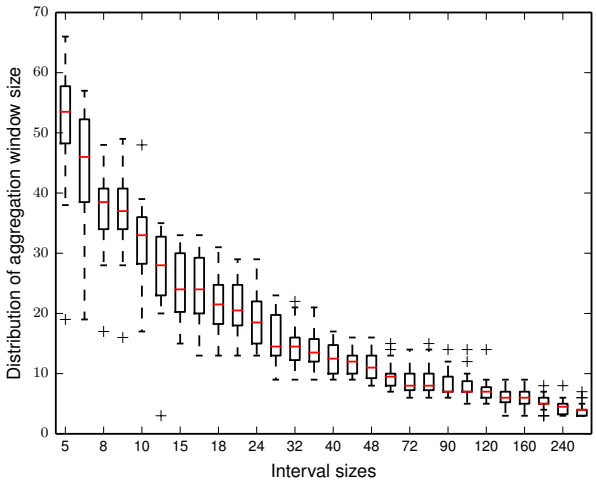


Figure 7.7: Distribution of aggregation window size vs. interval sizes.

7.8.4 Aggregation window Size

In the next experiment, we determine the minimum length of an aggregation window that is required to calculate a stable aggregated activity curve for our smart home data. Figure 7.7 shows the variations in the length of an aggregate window at different interval sizes calculated using all the available sensor data. We observe that the length of the aggregation window is larger for the smaller interval sizes and smaller for the larger interval sizes. We can explain such differences in length of the aggregation window based on the observations we made between average pairwise distances and interval sizes in Figure 7.7. At larger interval sizes, activity distributions are dominated by activities that take a long time to complete (such as sleep). Thus, the distance between two activity distributions for such activity curves are significantly lower than the distance between two activity distributions for activity curves at smaller time intervals. Hence, we obtain a stable activity curve using a smaller aggregation window size for larger interval sizes.

7.8.5 Change scores and correlations

In this section, we study the strength of the correlations between the changes detected in activity routines by the PCAR algorithm and the corresponding standard clinical scores (RBANS and TUG) for a smart home resident. Specifically, we

calculate correlations between change scores calculated by applying PCAR on activity curves derived using activity-labelled smart home sensor data and corresponding clinical scores ensuring that each pair of the smart home change score and clinical score was observed at around the same time.

To obtain such correlations, we first calculate aggregated activity curves for two three-month aggregation windows, W_1 and W_2. Next, we apply PCAR to these activity curves to obtain a smart home-based change score. We also obtain clinical scores measured at time points, t_1 and t_2. We repeat this step for all available pairs of consecutive testing time points for all 18 residents. Finally, we calculate Pearson correlation and Spearman rank correlation between the activity change scores and the corresponding clinical scores to evaluate the strength of the relationship. The process is summarized in Algorithm 7.

To evaluate our automated health correlation based on smart home data, we derived correlation coefficients between change scores obtained from the smart home-based activity curve model with the standard health clinical scores (TUG and RBANS scores). To conduct this experiment, we ran 1500 permutation iterations and derived change scores for both alignment techniques. We repeated the experiments for different time interval sizes.

As a baseline for comparison, we generated random change scores by randomly predicting a change between activity distributions instead of using the PCAR algo-

Algorithm 7 BehaviorAndHealthCorrelation(t)

1: // t = testing time points

2: // Return correlation coefficient

3: BehaviorChangeScores = [ ]

4: ClinicalChangeScores = [ ]

5: i = 0

6: **repeat**

7: Σ_i = AggregateActivityCurvesAtTime(t_i + 3 months)

8: Σ_{i+1} = AggregateActivityCurvesAtTime(t_{i+1} − 3 months)

9: $\hat{Dist}$ = EmpiricalDistribution($\Sigma_i, \Sigma_{i+1}, N_P$)

10: S_1 = PCAR ($\Sigma_i, \Sigma_{i+1}, \hat{Dist}$)

11: S_2 = ClinicalScores(t_i, t_{i+1})

12: Append(S_1 , BehaviorChangesScores)

13: Append(S_2, ClinicalChangesScores)

14: $i = i + 1$

15: **until**

16: **return** Correlation(S_1, S_2)

Table 7.2: Pearson (r) and Spearman rank (rho) correlations between activity change scores and RBANS scores.

	Time interval change score		DTW score		Random scores	
intervalsize	r-RBANS	rho-RBANS	r-RBANS	rho-RBANS	r-RBANS	rho-RBANS
5	0.00	-0.10	0.11	-0.04	0.10	0.06
6	-0.01	-0.13	0.06	-0.10	0.14	0.14
8	-0.01	-0.16	0.04	-0.08	-0.06	0.00
9	-0.03	-0.19	0.03	-0.09	0.09	0.11
10	-0.03	-0.15	-0.03	-0.13	-0.07	-0.10
12	-0.04	-0.17	-0.03	-0.13	-0.07	-0.07
15	-0.04	-0.14	-0.05	-0.16	0.11	0.26
16	-0.04	-0.13	-0.04	-0.14	-0.19	-0.19
18	-0.06	-0.18	-0.08	-0.13	0.02	0.04
20	-0.04	-0.18	-0.03	-0.12	-0.16	-0.13
24	-0.04	-0.18	-0.03	-0.16	0.05	0.04
30	-0.07	-0.20	-0.01	-0.17	-0.06	-0.11
32	-0.04	-0.20	0.01	-0.19	0.20	0.19
36	-0.04	-0.21	0.01	-0.17	0.06	0.01

rithm. Table 7.2 and 7.3 list the correlations between these two scores for different time interval sizes.

We make the following observations:

- We obtain statistically significant correlations between activity change scores and TUG scores (Table 7.3)
- No correlations exist between activity change scores obtained from random predictions and TUG scores.
- No correlations exist between smart home based activity change scores and RBANS scores.
- Often, the strength of correlations at larger time interval sizes is weak because at larger time intervals activities are either dominated by sleep activity or other activity. Hence, changes in activity distributions at large time intervals are comparatively harder to detect.

In the next experiment, we use the Maximum Mean Discrepancy (MMD) distance measure to perform a two-sample test (Algorithm 6) to compare between activity distributions. The objective of this experiment is to determine the influence that particular distance measure has on the relationships between change scores and clinical scores. MMD measures a distance between two distributions when the distributions are represented as elements in Reproducing Kernel Hilbert space. We perform

Table 7.3: Pearson (r) and Spearman rank (rho) correlations between activity change scores and TUG scores (*$p < 0.05$, **$p < 0.005$).

	Time interval index score		DTW score		Random scores	
intervalsize	r-TUG	rho-TUG	r-TUG	rho-TUG	r-TUG	rho-TUG
5	0.27	0.28	0.27	0.43**	-0.21	-0.06
6	0.28*	0.31*	0.27	0.40**	-0.09	-0.03
8	0.32*	0.39**	0.33*	0.37*	-0.09	-0.12
9	0.31*	0.35*	0.24	0.25	-0.15	-0.06
10	0.33*	0.31*	0.29*	0.37*	0.10	0.12
12	0.35*	0.33*	0.30*	0.30*	0.18	0.15
15	0.33*	0.34*	0.31*	0.40**	0.06	-0.06
16	0.34*	0.38*	0.30*	0.35*	-0.03	-0.10
18	0.32*	0.29*	0.29*	0.27	-0.17	-0.07
20	0.32*	0.35*	0.29*	0.28*	0.19	0.12
24	0.35*	0.29*	0.31*	0.23	-0.10	-0.13
30	0.33*	0.24	0.28	0.27	-0.15	-0.15
32	0.33*	0.28*	0.28*	0.32*	-0.15	-0.13
36	0.32*	0.28*	0.28	0.31*	0.01	0.03

a two-sample test to test the null hypothesis that two activity distributions $\mathbf{D}_{1,t}$ and $\mathbf{D}_{2,t}$ for time interval t are obtained from the same distribution against the alternative that $\mathbf{D}_{1,t}$ and $\mathbf{D}_{2,t}$ are obtained from different distributions. Given n training examples from $\mathbf{D}_{1,t}$, m samples from $\mathbf{D}_{2,t}$, and a kernel function k, the MMD can be empirically estimated as [58]:

$$MMD(\mathbf{D}_{1,t}, \mathbf{D}_{2,t}) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} [k(\mathbf{D}_{1,i,t}, \mathbf{D}_{1,j,t}) + k(\mathbf{D}_{2,i,t}, \mathbf{D}_{2,j,t}) - k(\mathbf{D}_{1,i,t}, \mathbf{D}_{2,j,t})] \tag{7.6}$$

Using MMD as a distance metric, we repeat all the steps in Algorithm 7 to calculate the correlations between change scores and clinical scores. While Tables 7.2 and 7.3 list the correlations between KL-divergence-based change scores and clinical scores, Table 7.4 lists the correlations between the change scores calculated by using the MMD-based two-sample test and clinical scores at different window sizes. We note that we calculate the MMD-based change scores using the time interval index-based method (see Section 7.5.1). While we observe significant correlations between TUG and MMD-based change scores, we do not observe significant correlations between change scores and the RBANS clinical scores. This observation about the strength of the correlations between MMD-based change scores and clinical scores are similar to correlations when we use the KL-divergence-based change scores.

Table 7.4: Pearson (r) and Spearman rank (rho) correlations between activity change scores calculated by using MMD-based two-sample test and RBANS and TUG scores (*$p < 0.05$, **$p < 0.005$).

	Time interval index score			
intervalsize	r-RBANS	r-TUG	rho-RBANS	rho-TUG
5	0.15	0.23	-0.04	0.36*
6	0.14	0.24	-0.06	0.34*
8	0.12	0.26	-0.08	0.38*
9	0.11	0.28*	-0.12	0.43**
10	0.08	0.24	-0.14	0.30*
12	0.09	0.28*	-0.14	0.35*
15	0.12	0.22	-0.08	0.30*
16	0.08	0.25	-0.14	0.36*
18	0.11	0.23	-0.07	0.26
20	0.11	0.20	-0.06	0.22
24	0.13	0.20	-0.08	0.22
30	0.14	0.15	-0.09	0.19
32	0.20	0.16	0.00	0.20
36	0.16	0.16	-0.08	0.24

7.8.6 Continuous change scores

In the previous section, we predicted change scores at six month intervals to correlate the smart home-based behavior change scores with changes in standard clinical scores. We can also calculate these change scores more frequently by running the PCAR algorithm on the activity curves that lies within a sliding window of size six months and shifting this sliding window by one month (30 days). We will refer to such frequent change scores as continuous change scores. We can use continuous change scores to monitor the "performance" of a smart home resident's everyday behavior.

Figure 7.8 shows how the continuous change scores of two smart home residents have varied with time. Each point in a plot represents a total change score obtained by using the PCAR algorithm to compare behavior six months prior with current behavior. First, we plot continuous change scores of a resident whose health status has declined (Figure 7.8). We observe that after a year, the total change score of this resident started to fluctuate. Such fluctuations indicate changes in the average daily routines of this resident. Similarly, we plot continuous change scores of another resident whose health has been in excellent condition for the entire data collection period (Figure 7.8). We observe that the PCAR algorithm detects very few changes in the average daily routines of this resident.

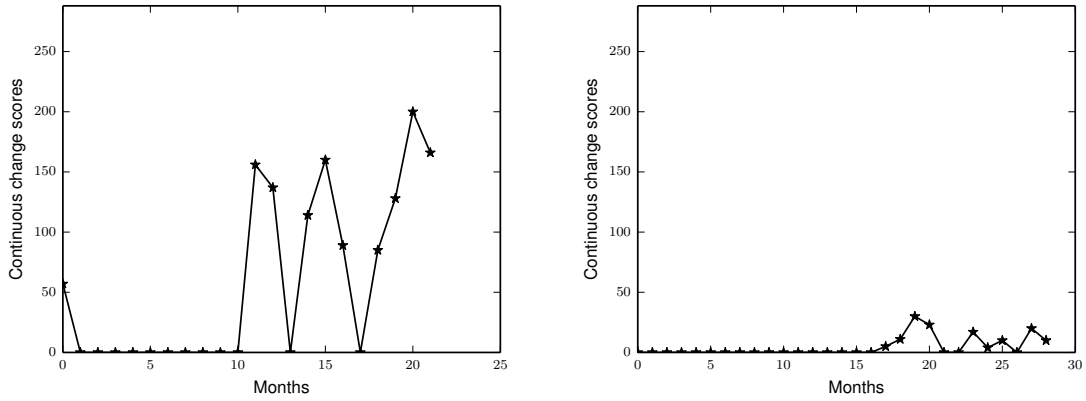


Figure 7.8: The continuous change scores of two residents calculated by running PCAR algorithm on a sliding window of six months with an aggregation window size of 30 days.

7.8.7 Individual activity change

The change scores calculated in the previous sections quantify overall changes in average daily routines for the entire collection of know activities. In this experiment, we can quantify total changes in the average daily routines of some specific activities by running PCAR algorithm on a reduced activity set. The elements in this reduced activity set are activities that we want to monitor and the "other activities" class is used to represent all of the remaining activities. For example, if we want to monitor sleep and bed to toilet activities, we put three elements (sleep, bed to toilet and other) in the reduced activity set. Using this reduced set of activity, we can use PCAR to

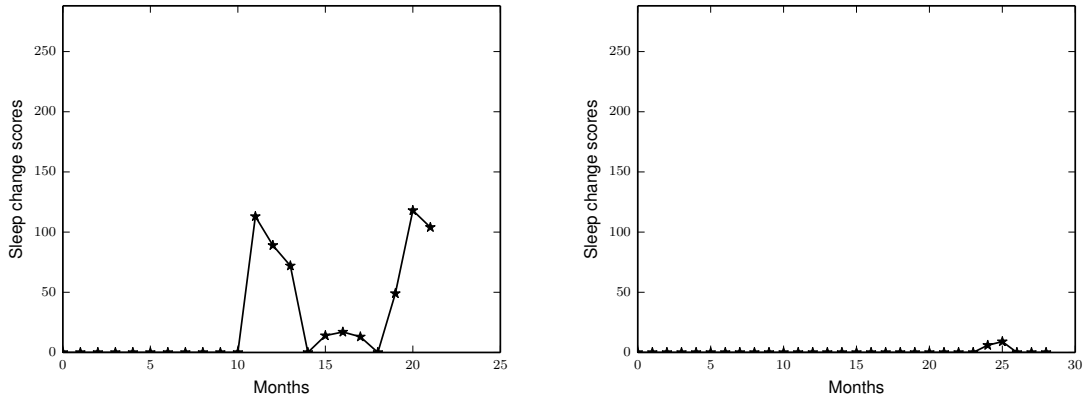


Figure 7.9: The continuous sleep change scores of two residents calculated by running PCAR algorithm on a sliding window of six months with an aggregation window size of $x = 30$ days.

obtain continuous change scores.

Figure 7.9 shows the continuous sleep change scores of the same two smart home residents for whom we studied the continuous change scores in Figure 7.8. We see that PCAR detected changes in the overall sleep routine of the first resident while it does not detect any sleep routine changes for the other resident.

7.9 Discussion and observations

In this chapter, we proposed an activity curve model to represent daily activity-based behavior routines. The proposed activity curve models the activity distribu-

tions of the activities at different times of a day. Using the activity curve model, we developed the PCAR algorithm to identify and quantify changes in the activity routines. We validated our model by performing experiments using synthetic data and longitudinal smart home sensor data. PCAR is able to represent behavior patterns through big data analysis.

The current activity model considers activity distributions using different interval sizes. In the future, we will build a hierarchical activity curve model to combine the activity distributions at different time interval sizes. We will also investigate techniques to extend the activity curve algorithm to detect acute health care events such as falls by using a shorter aggregated window size (such as a day or a week). Further, while performing experiments with an activity curve model, we choose a subset of activities that are considered important in daily life. In the future, we will extend our experiments to include a larger pool of daily activities. We also note that the activities that did not fit into these seven predefined categories were termed "other". We note that the "other" data is very large, complex, and represents important activities that we will add to our activity vocabulary in future work. A major contribution of this work is the introduction of an activity curve, which is a probabilistic representation of an aggregated daily routine. In our work, we validated the activity curve using environment sensors. We can also obtain activity curve representation using sensor data collected from other sources such as the sensor data from smart phones

and wearable devices. In the future, we will conduct further studies to explore how clinicians and caregivers can benefit from the activity curve representation and the changes in the daily routine that the algorithm detects.

We developed the PCAR algorithm to quantify changes in an activity routine. PCAR makes use of a smart home sensor data of an individual collected over a period to quantify changes in the activity routine and outputs change scores. This algorithmic approach is important because activity routines vary among individuals.

Furthermore, we studied the relationship between the output from the PCAR algorithm and the standard clinical and physical health scores. We found moderate correlations between the change scores and standard TUG scores. However, we found that the correlations between smart home-based change scores and standard cognitive scores (RBANS) were not as strong as we expected because the majority of the older adults for whom we analyzed the data are healthy older adults. Similarly, we also demonstrated methods to evaluate the "average performance" of a smart home resident by continuously monitoring changes in the overall daily routine as well as a set of specified activities. This chapter confirms that pervasive computing methods can be used to correlate an individual's behavior patterns and clinical assessment scores.

Unlike Chapters 4, 5 and 6 in which we model the performance of specific activities, in this chapter we propose an activity curve to model the daily routine

of a smart home resident using longitudinal smart home sensor data. We developed the PCAR algorithm to detect changes in the daily routine and use this change information to analyze the possibility of changes in the cognitive or physical health of a smart home resident.

CHAPTER 8. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

In this dissertation, we explored methods to utilize sensor data to model the everyday behavior of a smart home resident and apply machine learning algorithms to predict their well being. To validate the proposed algorithms, we utilized cross-sectional and longitudinal data collected from different smart home-based studies.

We note that the nature of data that we collect and the type of questions we answer are different in longitudinal and cross-sectional studies. We perform cross-sectional studies by observing a population at a single time point. Cross-sectional studies allow researchers to compare variables of different population groups. For example, a cross-sectional study allows us to answer if the task quality of the MCI group is different from the cognitively healthy group. In contrast, longitudinal studies are conducted by measuring observations of the same individual usually over a long period of time. The data from longitudinal studies allows researchers to answer questions related to within-individual changes and inter-individual differences in changes, trends, and trajectories over time. For example, a longitudinal study allows researchers to answer a question such as how the task quality of a particular individual has changed over time.

We first introduced a machine learning-based method for assessing activity quality in smart homes by utilizing the data collected from a cross-sectional smart home-based study. In this study, we recruited participants to perform a set of simple and complex activities. The participants performed simple daily activities such as eating and cooking. They also performed a complex activity named the Day Out Task in the smart home. Utilizing the sensor data that is collected while participants were performing activities, we developed learning algorithms to automatically predict the activity task quality. We also assessed the ability of learning algorithms to predict the cognitive health of the participants based on a task quality measure. Our results suggest that it is possible to automatically quantify the task quality of the smart home activities and perform limited assessment of the cognitive health of individuals by properly choosing smart home activities and training learning algorithms.

Using the longitudinal smart home sensor data, we developed algorithms to monitor the daily behavior of a smart home resident. The longitudinal data is collected for more than two years from 18 smart homes without interrupting or manipulating the residents' daily routine. We first introduced the Clinical Assessment using Activity Behavior (CAAB) algorithm to predict the cognitive and mobility scores of smart home residents by monitoring a set of basic and instrumental activities of daily living. We evaluate the performance of CAAB utilizing smart home sensor data collected from 18 smart homes over two years using prediction and classification-based

experiments. Our prediction and classification-based results suggest that it is feasible to predict standard clinical scores using smart home sensor data and learning-based data analysis.

Finally, we introduce an activity curve to represent an abstraction of an individual's daily behavioral routines. We propose methods to detect changes in daily behavioral routines by comparing activity curves and use these changes to analyze the possibility of changes in cognitive or physical health. We evaluate our change detection approach using a longitudinal smart home sensor dataset collected from smart homes with older adult residents. We demonstrate how activity curve-based change detection can be used to perform functional health assessment and our evaluation indicates that correlations do exist between behavior and health changes and that these changes can be automatically detected.

Future research directions

We identify three possible avenues for future work. First, we note that we performed longitudinal studies utilizing the data from 18 smart home residents. As a part of future work, we want to validate the proposed algorithms using a larger pool of population sizes encompassing a greater period. Additionally, our experiments rely on set of basic and instrumental activities of daily living. In the future, we would like

to experiment using wide variety of both shorter and longer-duration activities.

We will also explore the clinical utility of smart home-based predictions and the role it can play in helping clinicians to make informed decisions. Clinical data points are often sparse because clinicians administer clinical assessments on patients between wider intervals. In our smart home-based longitudinal study, the clinical tests were administered once in every six months. We can augment such sparse clinical datasets with the smart home-based predictions. We will answer the question of how clinicians can benefit from these predictions by developing visualization tools and carrying out additional studies [82].

Finally, we would like to extend the proposed algorithms to develop algorithms that detect early indications of cognitive and physical decline. The algorithm requires smart home longitudinal data in which there are known instances of health decline. Such algorithms will play pivotal roles in prevention and early intervention of cognitive and physical decline.

We envision this ongoing research as a critical component with smart home functionality that constantly tracks the behavior of its residents, automatically assesses their health, and provides an environment to senior residents where they can live independently and safely in their own homes.

APPENDIX

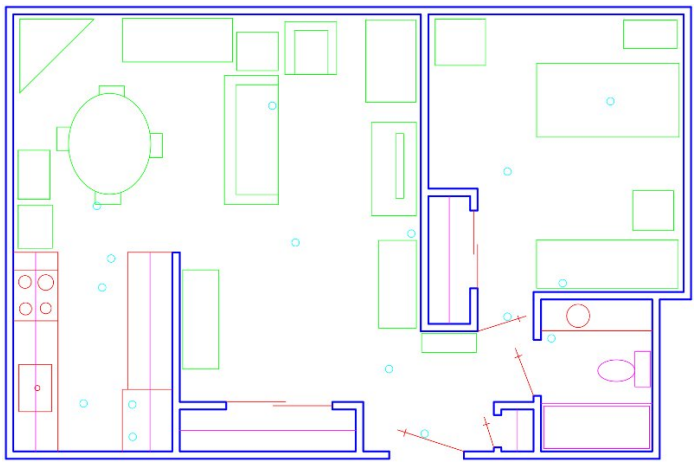
A Pervasive technological approaches to monitoring everyday functioning measures

Everday Measures	Studies	Technologies
Computerized Cognitive Assessment		
CERAD,CANTAB CANTAB Mobile	[48],[141]	
ClockMe	[77]	Desktop Computers,
Virtual Reality	[166]	Tablets, Smart Phones
Computer Usage		
Typing Speed	[72]	
Mouse Usage		
Computer Games		
Solitaire	[71], [70],[73]	
Word Scramble		
Mobility and Gait		
Gait Velocity	[11],[12],[61]	Motion Sensors
Everyday Functioning		
Object Usage	[63]	RFID
Simple Activities	[34],[41],[40]	Motion Sensors
Complex Activities	[59][60]	Motion Sensors
Longitudinal Data		
Visualization Tools	[143]	Motion Sensors
Statistical Models	[45][146]	

B Sensor dataset details for the CASAS longitudinal testbeds

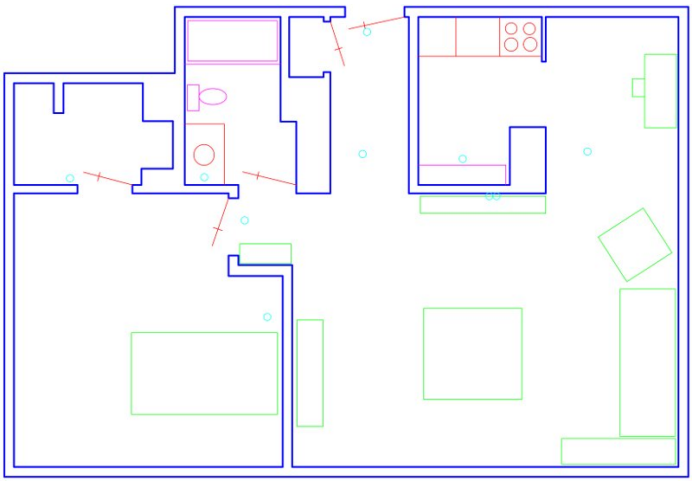
Apartment Id	Start date	End date	#Events
A	2012-07-18	2013-07-25	1895875
B	2011-06-13	2014-03-31	6496858
C	2011-06-10	2014-03-31	3608023
D	2011-06-09	2013-09-04	6607015
E	2011-06-09	2014-01-12	3875166
F	2011-06-09	2014-03-10	4418547
G	2011-06-10	2014-03-31	5875856
H	2011-06-10	2014-03-31	8917824
I	2011-06-10	2014-03-31	5505137
J	2011-06-09	2012-11-05	3262762
K	2011-06-10	2014-01-12	4698358
L	2011-06-10	2012-05-25	2263036
M	2011-06-13	2012-05-25	1131009
N	2011-06-10	2014-03-31	7463753
O	2012-01-27	2014-03-31	2612377
P	2013-02-28	2013-09-29	1332811
Q	2013-03-01	2014-03-31	1729474
R	2013-02-28	2014-03-31	1875665

C Floor plans for the CASAS longitudinal testbeds

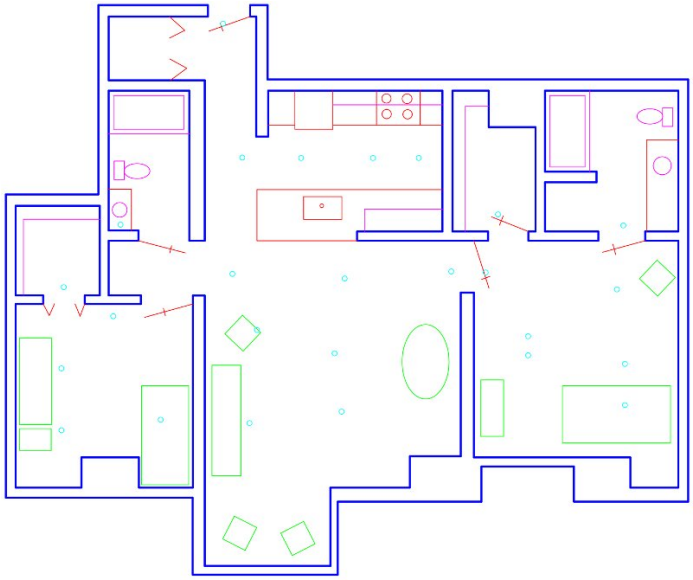


Apartment A

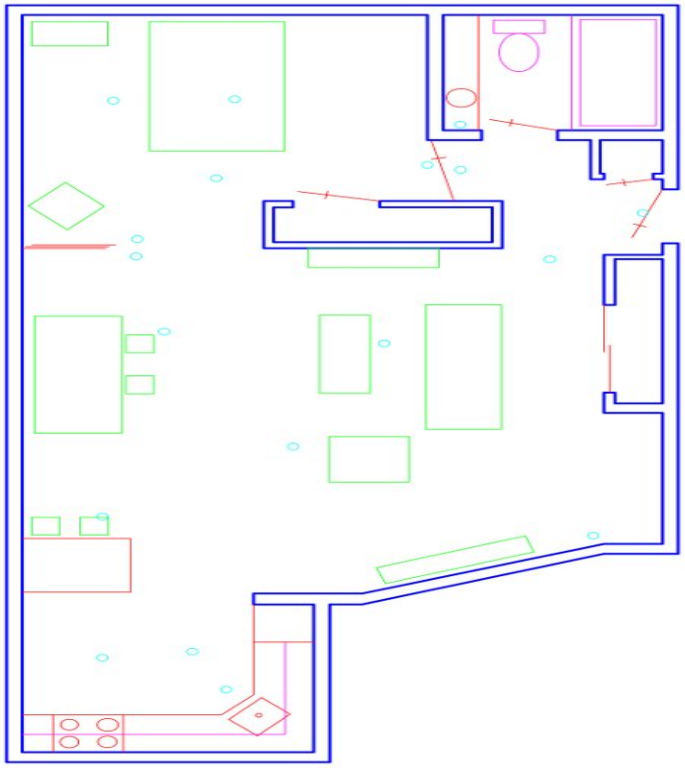
Apartment B



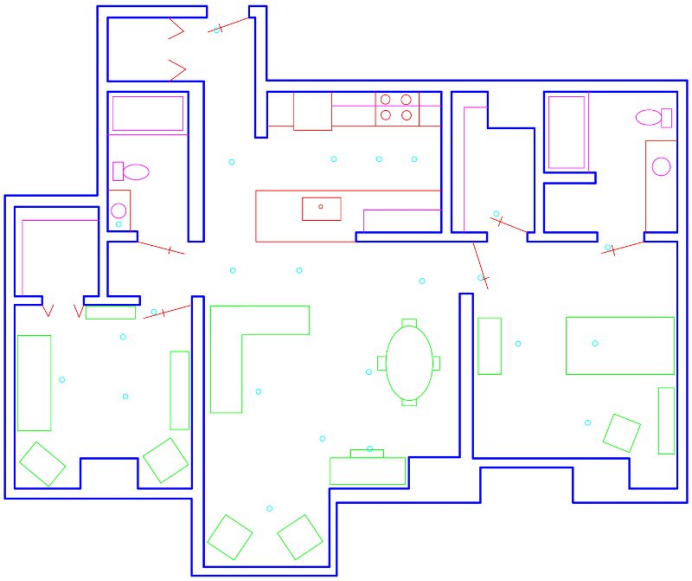
Apartment C



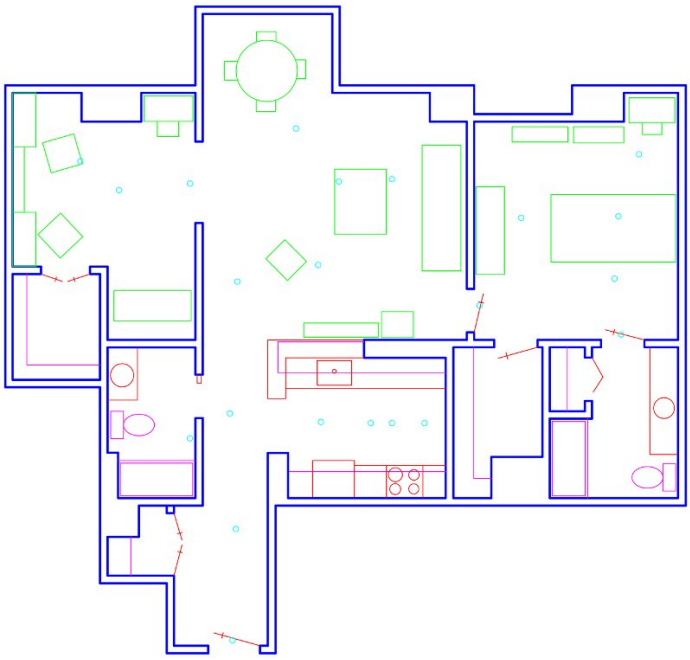
Apartment D



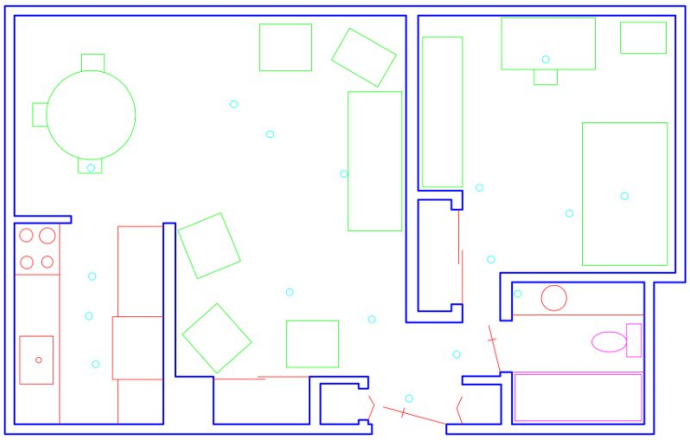
Apartment E



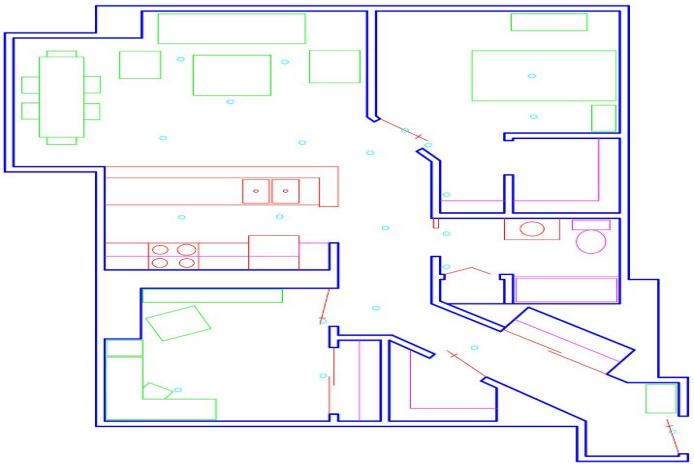
Apartment F



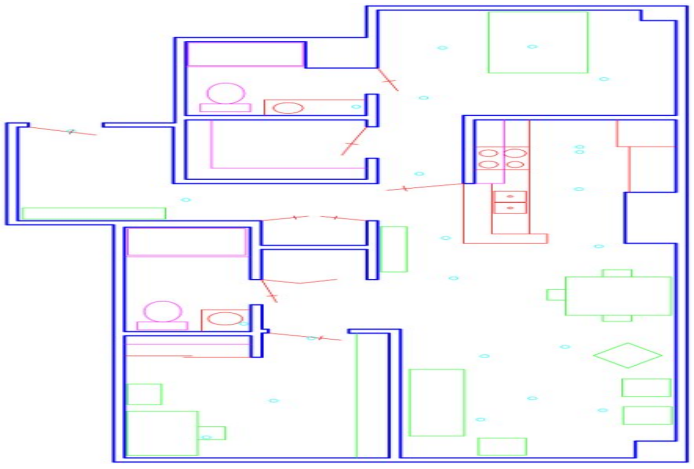
Apartment G



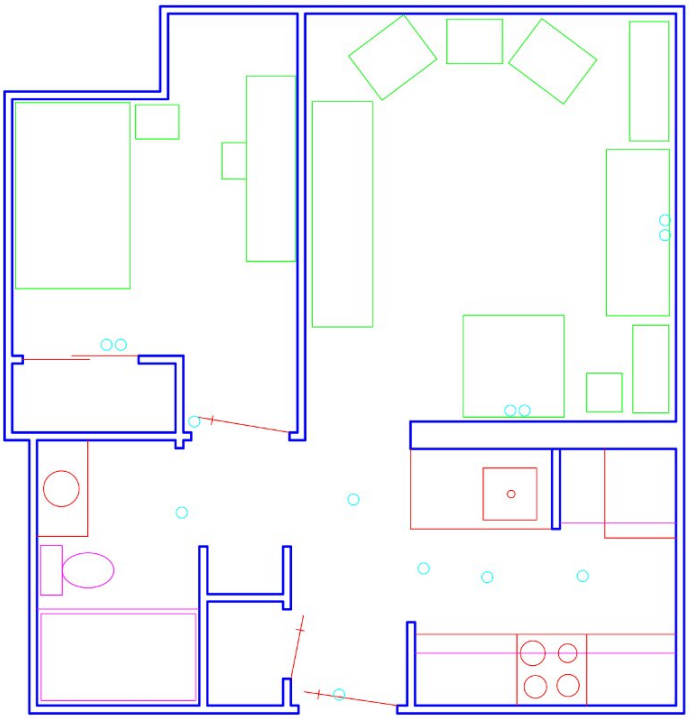
Apartment H



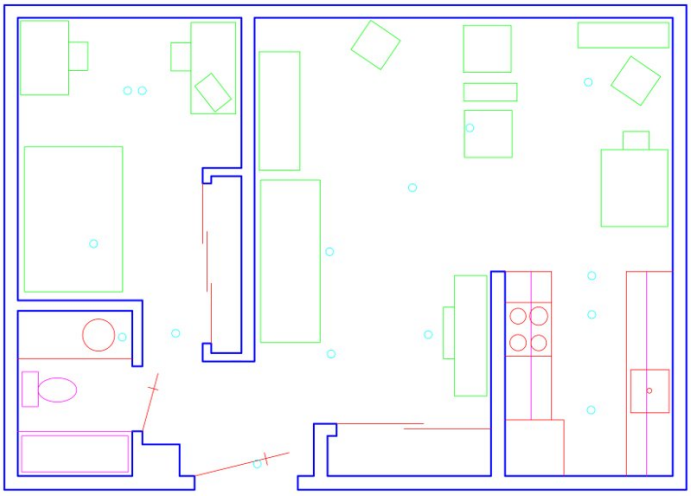
Apartment I



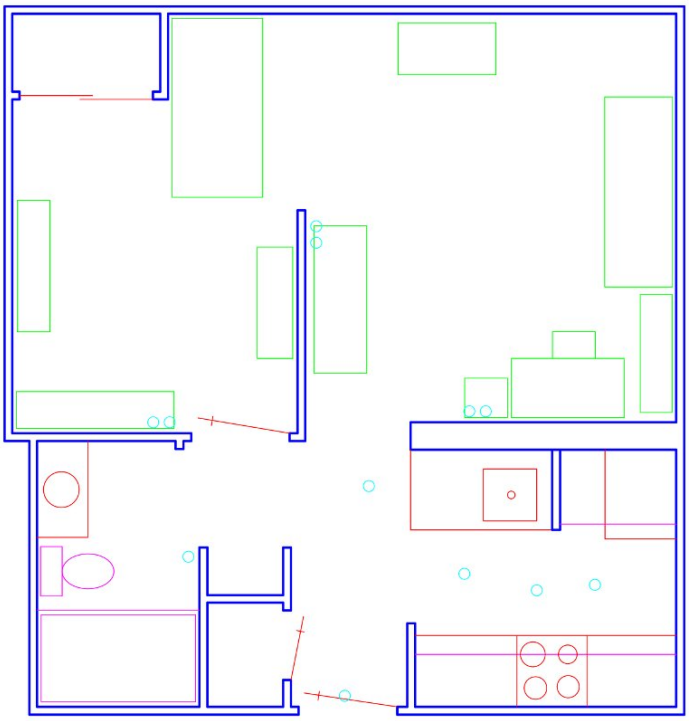
Apartment J



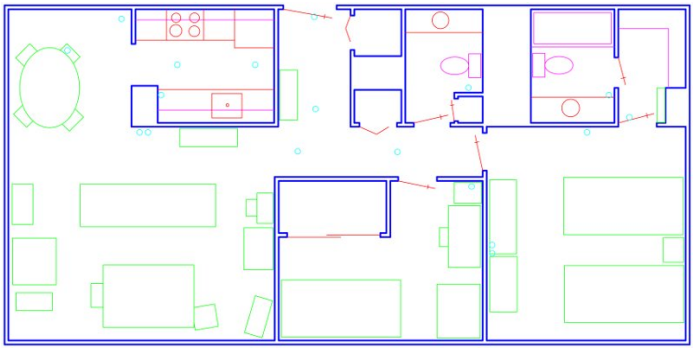
Apartment K



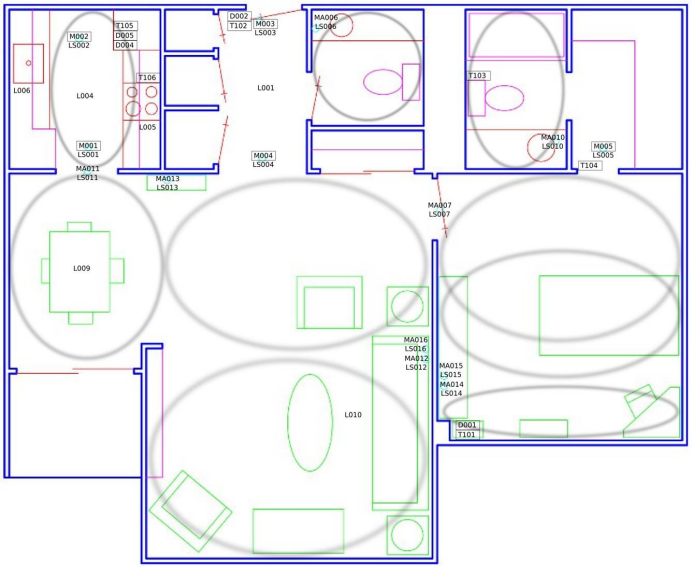
Apartment L



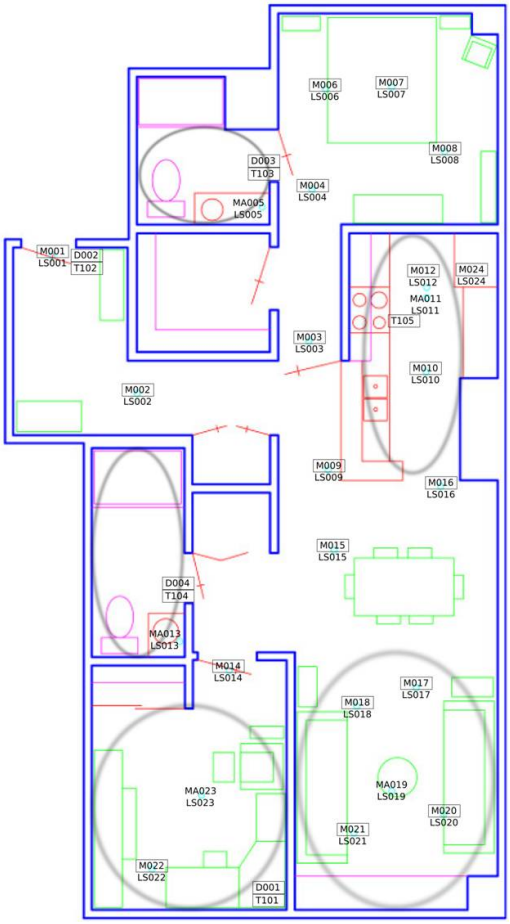
Apartment M



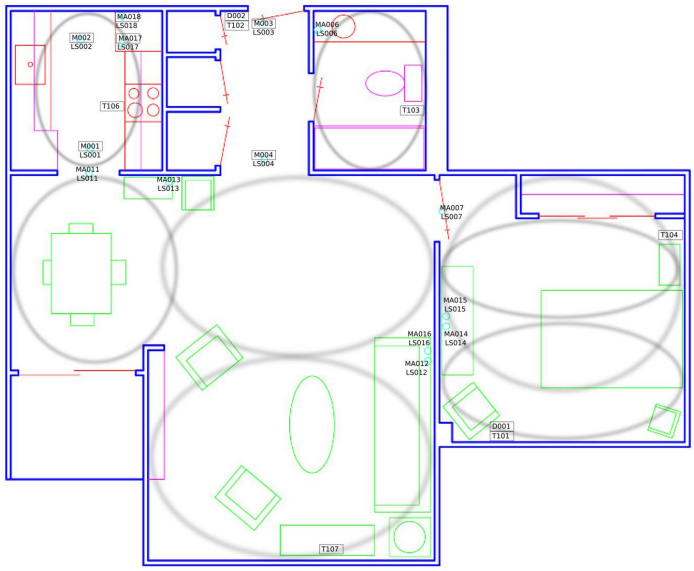
Apartment N



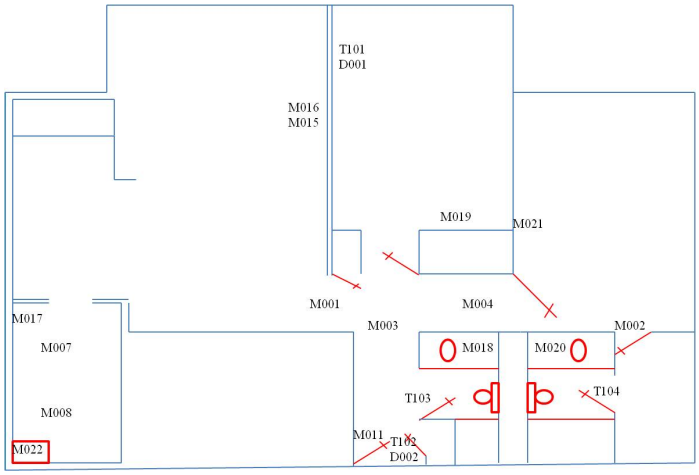
Apartment O



Apartment P



Apartment Q



Apartment R

REFERENCES

[1] Stopping elderly accidents, deaths & injuries. Center for Disease Control and Prevention.

[2] R. Aarhus, E. Grönvall, S.B. Larsen, and S. Wollsen. Turning training into play: Embodied gaming, seniors, physical training and motivation. *Gerontechnology*, 10(2):110–120, April 2011.

[3] J.K. Aggarwal and M.S. Ryoo. Human activity analysis. *ACM Computing Surveys*, 43(3):1–43, April 2011.

[4] J. Aitchison, C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn. Logratio Analysis and Compositional Distance. *Mathematical Geology*, 32(3):271–275, 2000.

[5] Marilyn S. Albert, Steven T. DeKosky, Dennis Dickson, Bruno Dubois, Howard H. Feldman, Nick C Fox, Anthony Gamst, David M. Holtzman, William J. Jagust, Ronald C. Petersen, Peter J Snyder, Maria C Carrillo, Bill Thies, and Creighton H Phelps. 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 : the journal of the Alzheimer's Association*, 7(3):270–9, May 2011.

[6] Association Alzheimer's. 2012 Alzheimer's disease facts and figures. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 8(2):131 – 168, 2012.

[7] O. Amft and G. Troster. On-Body Sensing Solutions for Automatic Dietary Monitoring. *IEEE Pervasive Computing*, 8(2):62–70, April 2009.

[8] Oliver Amft. Self-Taught Learning for Activity Spotting in On-body Motion Sensor Data. In *2011 15th Annual International Symposium on Wearable Computers*, pages 83–86. IEEE, June 2011.

[9] Artero S., Touchon J., and Ritchie K. Disability and mild cognitive impairment: a longitudinal populationbased study. *International Journal of Geriatric Psychiatry*, 16(11):1092–1097, 2001.

[10] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-IV-TR.*, volume 4th of *Diagnostic and statistical manual*

of mental disorders. American Psychiatric Association, Washington, DC, 4 edition, 2000.

[11] Daniel Austin, Tamara L Hayes, Jeffrey Kaye, Nora Mattek, and Misha Pavel. On the Disambiguation of Passively Measured In-home Gait Velocities from Multi-person Smart Homes. *Journal of ambient intelligence and smart environments*, 3(2):165–174, January 2011.

[12] Daniel Austin, Tamara L Hayes, Jeffrey Kaye, Nora Mattek, and Misha Pavel. Unobtrusive monitoring of the longitudinal evolution of in-home gait velocity data with applications to elder care. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 2011, pages 6495–6498, January 2011.

[13] Daniel Austin, Todd Leen, Tamara Hayes, Jeff Kaye, Holly Jimison, Nora Mattek, and Misha Pavel. Model-based inference of cognitive processes from unobtrusive gait velocity measurements. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 2010, pages 5230–5233, Buenos Aires, Argentina, January 2010.

[14] RHB Benedict. *Brief Visuospatial Memory Test Revised.* Psychological Assessment Resources, Odessa, FL, 1997.

[15] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289 – 300, 1995.

[16] Ulf Blanke, Bernt Schiele, Matthias Kreil, Paul Lukowicz, Bernhard Sick, and Thiemo Gruber. All for one or one for all? Combining heterogeneous features for activity spotting. In *2010 8th IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 18–24. IEEE, March 2010.

[17] Alesandro Ble, Stefano Volpato, Giovanni Zuliani, Jack M Guralnik, Stefania Bandinelli, Fulvio Lauretani, Benedetta Bartali, Cinzia Maraldi, Renato Fellin, and Luigi Ferrucci. Executive function correlates with walking speed in older persons: the InCHIANTI study. *Journal of the American Geriatrics Society*, 53(3):410–5, March 2005.

[18] J. Brandt, M. Spencer, and M. Folstein. The telephone interview for cognitive status. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, 1:111–117, 1988.

[19] Michael Buettner, Richa Prasad, Matthai Philipose, and David Wetherall. Recognizing daily activities with RFID-based sensors. In *Proceedings of the 11th international conference on Ubiquitous computing - Ubicomp '09*, page 51, New York, New York, USA, September 2009. ACM Press.

[20] Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys*, 46(3):1–33, January 2014.

[21] Andreas Bulling, Jamie A. Ward, and Hans Gellersen. Multimodal recognition of reading activity in transit using body-worn sensors. *ACM Transactions on Applied Perception*, 9(1):1–21, March 2012.

[22] P W Burgess. Strategy application disorder: the role of the frontal lobes in human multitasking. *Psychological research*, 63(3-4):279–88, January 2000.

[23] Naomi Chaytor and Maureen Schmitter-Edgecombe. The ecological validity of neuropsychological tests: a review of the literature on everyday cognitive skills. *Neuropsychology review*, 13(4):181–197, December 2003.

[24] Naomi Chaytor, Maureen Schmitter-Edgecombe, and Robert Burr. Improving the ecological validity of executive functioning assessment. *Archives of clinical neuropsychology*, 21(3):217–27, April 2006.

[25] Chao Chen, Diane J. Cook, and Aaron S. Crandall. The user side of sustainability: Modeling behavior and energy usage in the home. *Pervasive and Mobile Computing*, 9(1):161–175, 2013.

[26] Liming Chen, J. Hoey, C.D. Nugent, and Diane Cook. Sensor-Based Activity Recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):790–808, November 2012.

[27] Diane J. Cook. Learning Setting-Generalized Activity Models for Smart Spaces. *IEEE intelligent systems*, 27(1):32–38, September 2010.

[28] Diane J Cook. Computer science. How smart is your home? *Science (New York, N.Y.)*, 335(6076):1579–81, March 2012.

[29] Diane J. Cook, Aaron S. Crandall, Brian L. Thomas, and Narayanan C. Krishnan. CASAS: A Smart Home in a Box. *Computer*, 46(7):62–69, July 2013.

[30] Diane J. Cook and Sajal Das. *Smart Environments: Technology, Protocols and Applications (Wiley Series on Parallel and Distributed Computing)*. Wiley-Interscience, October 2004.

[31] Diane J. Cook and Sajal K. Das. Pervasive computing at scale: Transforming the state of the art. *Pervasive and Mobile Computing*, 8(1):22–35, February 2012.

[32] Diane J. Cook and Narayanan C. Krishnan. *Activity Learning: Discovering, Recognizing, and Predicting Human Behavior from Sensor Data*. Wiley, New York, 2015.

[33] Diane J Cook, Narayanan C Krishnan, and Parisa Rashidi. Activity discovery and activity recognition: a new partnership. *Transactions on Systems, Man and Cybernetics, Part B*, 43(3):820–828, June 2013.

[34] Diane J. Cook and M. Schmitter-Edgecombe. Assessing the quality of activities in a smart environment. *Methods of Information in Medicine*, 48(5):480–485, January 2009.

[35] Antonia K Coppin, Anne Shumway-Cook, Jane S Saczynski, Kushang V Patel, Alessandro Ble, Luigi Ferrucci, and Jack M Guralnik. Association of executive function and performance of dual-task physical tests among older adults: analyses from the InChianti study. *Age and ageing*, 35(6):619–624, November 2006.

[36] Aaron S. Crandall. *Behaviometrics for multiple residents in a smart environment*. PhD thesis, Washington State University, 2011.

[37] Vasilis Dakos, Stephen R Carpenter, William A Brock, Aaron M Ellison, Vishwesha Guttal, Anthony R Ives, Sonia Kéfi, Valerie Livina, David A Seekell, Egbert H van Nes, and Marten Scheffer. Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. *PloS one*, 7(7):e41010, January 2012.

[38] Barnan Das. *Machine Learning Challenges for Automated Prompting in Smart Homes*. PhD thesis, Washington State University, 2014.

[39] Barnan Das, Diane J. Cook, Maureen Schmitter-Edgecombe, and Adriana M. Seelye. PUCK: an automated prompting system for smart environments: toward achieving automated promptingchallenges involved. *Personal and Ubiquitous Computing*, 16(7):859–873, September 2012.

[40] Prafulla Dawadi, Diane Cook, and Maureen Schmitter-Edgecombe. Automated cognitive health assessment using smart home smart monitoring of complex tasks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 43(6):1302–1313, 2013.

[41] Prafulla Dawadi, Diane J. Cook, Maureen Schmitter-Edgecombe, and Carolyn Parsey. Automated assessment of cognitive health using smart home technologies. *Technology and health care*, 21(4):323–343, April 2013.

[42] D.C. Delis. *Executive Function System: Examiners Manual.* The Psychological Corporation, San Antonio, Texas, 2001.

[43] Cynthia L Deschenes and Susan M McCurry. Current treatments for sleep disturbances in individuals with dementia. *Current psychiatry reports*, 11(1):20–26, February 2009.

[44] Manfred Diehl, Michael Marsiske, Ann L Horgas, Adrienne Rosenberg, Jane S Saczynski, and Sherry L Willis. The Revised Observed Tasks of Daily Living: A Performance-Based Assessment of Everyday Problem Solving in Older Adults. *Journal of applied gerontology*, 24(3):211–230, 2005.

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

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

[47] X. Dong-Hui, A. S. Kurani, J. D. Furst, and D. S. Raicu. Run-length Encoding for Volumetric Texture. In *Proc. International Conference on Visualization Imaging and Image Processing*, pages 68–73, 2004.

[48] John H Dougherty, Rex L Cannon, Christopher R Nicholas, Lorin Hall, Felicia Hare, Erika Carr, Andrew Dougherty, Jennifer Janowitz, and Justin Arunthamakun. The computerized self test (CST): an interactive, internet accessible cognitive screening test for dementia. *Journal of Alzheimer's disease : JAD*, 20(1):185–195, January 2010.

[49] Henry C Driscoll, Linda Serody, Susan Patrick, Jennifer Maurer, Salem Bensasi, Patricia R Houck, Sati Mazumdar, Eric A Nofzinger, Bethany Bell, Robert D Nebes, Mark D Miller, and Charles F Reynolds. Sleeping well, aging well: a descriptive and cross-sectional study of sleep in "successful agers" 75 and older. *The American journal of geriatric psychiatry*, 16(1):74–82, January 2008.

[50] Saso Džeroski and Bernard Ženko. Is Combining Classifiers with Stacking Better than Selecting the Best One? *Machine Learning*, 54(3):255–273, March 2004.

[51] Eamonn Eeles. Sleep and its management in dementia. *Reviews in Clinical Gerontology*, 16(01):59–70, January 2007.

[52] Bradley Efron and R.J. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall, New York, 1994.

[53] Sarah T. Farias, Dan Mungas, Bruce R. Reed, Danielle Harvey, Deborah Cahn-Weiner, and Charles Decarli. MCI is associated with deficits in everyday functioning. *Alzheimer disease and associated disorders*, 20(4):217–223, 2006.

[54] Annette L Fitzpatrick, Catherine K Buchanan, Richard L Nahin, Steven T Dekosky, Hal H Atkinson, Michelle C Carlson, and Jeff D Williamson. Associations of gait speed and other measures of physical function with cognition in a healthy cohort of elderly persons. *The journals of gerontology. Series A, Biological sciences and medical sciences*, 62(11):1244–1251, November 2007.

[55] M F Folstein, S E Folstein, and P R McHugh. Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198, 1975.

[56] Kim Fouchenette. *Mobile Testing of Cognitive Function.* PhD thesis, Royal Instituite of Technology, 2010.

[57] Thomas Frenken and M Govercin. Precise assessment of self-selected gait velocity in domestic environments. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2010 4th International Conference*, pages 1–8, 2010.

[58] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–723–773–773, March 2012.

[59] Alden L Gross, George W Rebok, Frederick W Unverzagt, Sherry L Willis, and Jason Brandt. Cognitive Predictors of Everyday Functioning in Older Adults: Results From the ACTIVE Cognitive Intervention Trial. *The journals of gerontology Series B Psychological sciences and social sciences*, 66(5):557–566, 2011.

[60] J M Guralnik, L Ferrucci, C F Pieper, S G Leveille, K S Markides, G V Ostir, S Studenski, L F Berkman, and R B Wallace. Lower extremity function and subsequent disability: consistency across studies, predictive models, and value of gait speed alone compared with the short physical performance battery.

The journals of gerontology. Series A, Biological sciences and medical sciences, 55(4):221–231, April 2000.

[61] Stuart Hagler, Daniel Austin, Tamara L Hayes, Jeffrey Kaye, and Misha Pavel. Unobtrusive and ubiquitous in-home monitoring: a methodology for continuous assessment of gait velocity in elders. *IEEE transactions on bio-medical engineering*, 57(4):813–820, April 2010.

[62] C C Hoch, M A Dew, C F Reynolds, D J Buysse, P D Nowell, T H Monk, S Mazumdar, M D Borland, J Miewald, and D J Kupfer. Longitudinal changes in diary- and laboratory-based sleep measures in healthy "old old" and "young old" subjects: a three-year follow-up. *Sleep*, 20(3):192–202, March 1997.

[63] Mark R. Hodges, Ned L. Kirsch, Mark W. Newman, and Martha E. Pollack. Automatic assessment of cognitive impairment through electronic observation of object usage. In Patrik Floréen, Antonio Krüger, and Mirjana Spasojevic, editors, *Proc. International Conference on Pervasive Computing*, volume 6030 of *Lecture Notes in Computer Science*, pages 192–209, Berlin, Heidelberg, May 2010. Springer Berlin Heidelberg.

[64] Tracey Holsinger, Janie Deveau, Malaz Boustani, and John W Williams. Does this patient have dementia? *JAMA : The Journal of the American Medical Association*, 297(21):2391–2404, June 2007.

[65] T Hope, J Keene, K Gedling, C G Fairburn, and R Jacoby. Predictors of institutionalization for people with dementia living at home with a carer. *International journal of geriatric psychiatry*, 13(10):682–690, October 1998.

[66] Harold Hotelling. The Generalization of Student's Ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, August 1931.

[67] Robert J. Ivnik, James F. Malec, Glenn E. Smith, Eric G. Tangalos, and Ronald C. Petersen. Neuropsychological tests' norms above age 55: COWAT, BNT, MAE token, WRAT-R reading, AMNART, STROOP, TMT, and JLO. *The Clinical Neuropsychologist*, 10(3):262–278, July 1996.

[68] Angela L Jefferson, Robert H Paul, Al Ozonoff, and Ronald A Cohen. Evaluating elements of executive functioning as predictors of instrumental activities of daily living (IADLs). *Archives of clinical neuropsychology*, 21(4):311–320, May 2006.

[69] Marko Jelicic, Hans Bosma, Rudolf W H M Ponds, Martin P J Van Boxtel, Peter J Houx, and Jelle Jolles. Subjective sleep problems in later life as predictors of cognitive decline. Report from the Maastricht Ageing Study (MAAS). *International journal of geriatric psychiatry*, 17(1):73–77, January 2002.

[70] Holly Jimison and Misha Pavel. Embedded assessment algorithms within home-based cognitive computer game exercises for elders. In *International Conference of the IEEE Engineering in Medicine and Biology Society.*, volume 1, pages 6101–6104, New York, New York, USA, January 2006.

[71] Holly Jimison, Misha Pavel, and Thai Le. Home-based cognitive monitoring using embedded measures of verbal fluency in a computer word game. *International Conference of the IEEE Engineering in Medicine and Biology Society.*, 2008:3312–3315, January 2008.

[72] Holly Jimison, Misha Pavel, James McKanna, and Jesse Pavel. Unobtrusive monitoring of computer interactions to detect cognitive status in elders. *IEEE Engineering in Medicine and Biology Society*, 8(3):248–252, September 2004.

[73] Holly B. Jimison, Misha Pavel, Katherine Wild, Payton Bissell, James McKanna, Daniel Blaker, and Devin Williams. A Neural Informatics Approach to Cognitive Assessment and Monitoring. In *3rd International IEEE/EMBS Conference on Neural Engineering*, pages 696–699. IEEE, May 2007.

[74] Youn Joo Kang, Jeonghun Ku, Kiwan Han, Sun I Kim, Tae Won Yu, Jang Han Lee, and Chang Il Park. Development and clinical trial of virtual reality-based cognitive assessment in people with stroke: preliminary study. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, 11(3):3329–3339, June 2008.

[75] Marije Kanis, Saskia Robben, Judith Hagen, Anne Bimmerman, Natasja Wagelaar, and Ben Kröse. Sensor Monitoring in the Home : Giving Voice to Elderly People. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*, pages 97–100, Venice, Italy, 2013.

[76] T. L. M. Kasteren, G. Englebienne, and B. J. A. Kröse. An activity monitoring system for elderly care using generative and discriminative models. *Personal and Ubiquitous Computing*, 14(6):489–498, February 2010.

[77] Hyungsin Kim, CP Hsiao, and EYL Do. Home-based computerized cognitive assessment tool for dementia screening. *Journal of Ambient Intelligence and Smart Environments*, 4(5):429–442, 2012.

[78] Narayanan C Krishnan and Diane J Cook. Activity Recognition on Streaming Sensor Data. *Pervasive and mobile computing*, 10:138–154, February 2014.

[79] Narayanan C. Krishnan and Sethuraman Panchanathan. Analysis of low resolution accelerometer data for continuous human activity recognition. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3337–3340. IEEE, March 2008.

[80] Gaetan Lafortune and Gaëlle Balestat. Trends in Severe Disability Among Elderly People: Assessing the Evidence in 12 OECD Countries and the Future Implications. *OECD Health Working Papers*, March 2007.

[81] M. P. Lawton, M. Moss, M. Fulcomer, and M. H. Kleban. A research and service oriented multilevel assessment instrument. *Journal of gerontology*, 37(1):91–99, January 1982.

[82] Thai Le, Blaine Reeder, Daisy Yoo, Rafae Aziz, Hilaire J Thompson, and George Demiris. An evaluation of wellness assessment visualizations for older adults. *Telemedicine Journal and E-health*, 21(1):9–15, January 2015.

[83] Jonathan Lester, Tanzeem Choudhury, Nicky Kern, Gaetano Borriello, and Blake Hannaford. A hybrid discriminative/generative approach for modeling human activities. In *Proceedings of the 19th international joint conference on Artificial intelligence*, pages 766–772. Morgan Kaufmann Publishers Inc., July 2005.

[84] Jonathan Lester, Tanzeem Choudhury, Nicky Kern, Gaetano Borriello, and Blake Hannaford. A hybrid discriminative/generative approach for modeling human activities. pages 766–772, July 2005.

[85] Bao Ling and Stephen S. Intille. Activity Recognition from User-Annotated Acceleration Data. In *Pervasive, Lecture Notes in Computer Science*, pages 1–17, 2004.

[86] T Luck, M Luppa, M C Angermeyer, A Villringer, H-H König, and S G Riedel-Heller. Impact of impairment in instrumental activities of daily living and mild cognitive impairment on time to incident dementia: results of the Leipzig Longitudinal Study of the Aged. *Psychological medicine*, 41(5):1087–1097, May 2011.

[87] L. Maaten, E. Postma, and H. Herik. Dimensionality reduction: A comparative review. Technical report, Tilburg University Technical Report, 2009.

[88] Scott Marquis, M Milar Moore, Diane B Howieson, Gary Sexton, Haydeh Payami, Jeffrey A Kaye, and Richard Camicioli. Independent predictors of cognitive decline in healthy elderly persons. *Archives of neurology*, 59(4):601–606, April 2002.

[89] D. Marson and K. Hebert. Geriatric Neuropsychology Assessment and Intervention. In *Geriatric Neuropsychology Assessment and Intervention*, chapter Functional, pages 158–189. The Guilford Press, New York, USA, 2006.

[90] Jennifer L Martin, Lavinia Fiorentino, Stella Jouldjian, Karen R Josephson, and Cathy A Alessi. Sleep quality in residents of assisted living facilities: effect on quality of life, functional status, and depression. *Journal of the American Geriatrics Society*, 58(5):829–36, May 2010.

[91] K. Mase. Activity and location recognition using wearable sensors. *IEEE Pervasive Computing*, 1(3):24–32, July 2002.

[92] A Matic, P Mehta, J M Rehg, V Osmani, and O Mayora. Monitoring dressing activity failures through RFID and video. *Methods of information in medicine*, 51(1):45–54, January 2012.

[93] U. Maurer, A. Smailagic, D.P. Siewiorek, and M. Deisher. Activity Recognition and Monitoring Using Multiple Sensors on Different Body Positions. In *International Workshop on Wearable and Implantable Body Sensor Networks (BSN'06)*, pages 113–116. IEEE, 2006.

[94] Courtney McAlister and Maureen Schmitter-Edgecombe. Naturalistic assessment of executive function and everyday multitasking in healthy older adults. *Neuropsychology, development, and cognition. Section B, Aging, neuropsychology and cognition*, 20(6):735–56, January 2013.

[95] Giovanni Measso, Fabiano Cavarzeran, Giuseppe Zappalà, Barry D. Lebowitz, Thomas H. Crook, Francis J. Pirozzolo, Luigi A. Amaducci, Danilo Massari, and Francesco Grigoletto. The minimental state examination: Normative study of an Italian random sample. *Developmental Neuropsychology*, 9(2):77–85, April 1993.

[96] L. Middleton, A.A. Buss, A. Bazin, and M.S. Nixon. A Floor Sensor System for Gait Recognition. In *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*, pages 171–176, Buffalo, NY, USA, 2005. IEEE.

[97] J. C. Morris. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology*, 43(11):2412–2414, 1993.

[98] P J Murphy, N L Rogers, and S S Campbell. Age differences in the spontaneous termination of sleep. *Journal of sleep research*, 9(1):27–34, March 2000.

[99] Jan M Noyes and Kate J Garland. Computer- vs. paper-based tasks: are they equivalent? *Ergonomics*, 51(9):1352–1375, September 2008.

[100] Melissa L O'Connor, Jerri D Edwards, Virginia G Wadley, and Michael Crowe. Changes in mobility among older adults with psychometrically defined mild cognitive impairment. *The journals of gerontology. Series B, Psychological sciences and social sciences*, 65B(3):306–316, May 2010.

[101] Department of Health. Speech by the Rt Hon Patricia Hewitt MP, Secretary of State for Health, 2007.

[102] Georg Ogris, Thomas Stiefmeier, Paul Lukowicz, and Gerhard Troster. Using a complex multi-modal on-body sensor system for activity spotting. In *2008 12th IEEE International Symposium on Wearable Computers*, pages 55–62. IEEE, September 2008.

[103] Maurice M Ohayon, Mary A Carskadon, Christian Guilleminault, and Michael V Vitiello. Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: developing normative sleep values across the human lifespan. *Sleep*, 27(7):1255–1273, November 2004.

[104] Markus Ojala and Gemma C. Garriga. Permutation Tests for Studying Classifier Performance. *The Journal of Machine Learning Research*, 11:1833–1863, March 2010.

[105] Yoshitaka Ouchi, Kyoko Akanuma, Mitsue Meguro, Mari Kasai, Hiroshi Ishii, and Kenichi Meguro. Impaired instrumental activities of daily living affect conversion from mild cognitive impairment to dementia: the Osaki-Tajiri Project. *Psychogeriatrics*, 12(1):34–42, March 2012.

[106] Paula Paavilainen, Ilkka Korhonen, Jyrki Lötjönen, Luc Cluitmans, Marja Jylhä, Antti Särelä, and Markku Partinen. Circadian activity rhythm in demented and non-demented nursing-home residents measured by telemetric actigraphy. *Journal of sleep research*, 14(1):61–68, March 2005.

[107] Paula Paavilainen, Ilkka Korhonen, and Markku Partinen. Telemetric activity monitoring as an indicator of long-term changes in health and well-being of older people. *Gerontechnology*, 4(2):77–85, 2005.

[108] Paulito Palmes, Hung Keng Pung, Tao Gu, Wenwei Xue, and Shaxun Chen. Object relevance weight pattern mining for activity recognition and segmentation. *Pervasive and Mobile Computing*, 6(1):43–57, 2010.

[109] Misha Pavel, Tamara Hayes, Ishan Tsay, Deniz Erdogmus, Anindya Paul, Nicole Larimer, Holly Jimison, and John Nutt. Continuous Assessment of Gait Velocity in Parkinson's Disease from Unobtrusive Measurements. In *3rd International IEEE/EMBS Conference on Neural Engineering*, pages 700–703, Kohala Coast, Hawaii, May 2007. IEEE.

[110] H Pedrosa, A De Sa, M Guerreiro, J Maroco, M R Simoes, D Galasko, and A de Mendonca. Functional evaluation distinguishes MCI patients from healthy elderly people–the ADCS/MCI/ADL scale. *The journal of nutrition, health & aging*, 14(8):703–709, October 2010.

[111] K Pérès, V Chrysostome, C Fabrigoule, J M Orgogozo, J F Dartigues, and P Barberger-Gateau. Restriction in complex activities of daily living in MCI: impact on outcome. *Neurology*, 67(3):461–466, August 2006.

[112] Karine Pérès, Catherine Helmer, Hélène Amieva, Jean-Marc Orgogozo, Isabelle Rouch, Jean-François Dartigues, and Pascale Barberger-Gateau. Natural history of decline in instrumental activities of daily living performance over the 10 years preceding the clinical diagnosis of dementia: a prospective population-based study. *Journal of the American Geriatrics Society*, 56(1):37–44, January 2008.

[113] Robert Perneczky, Corina Pohl, Christian Sorg, Julia Hartmann, Natasa Tosic, Timo Grimmer, Sandra Heitele, and Alexander Kurz. Impairment of activities of daily living requiring memory or complex reasoning as part of the MCI syndrome. *International journal of geriatric psychiatry*, 21(2):158–62, February 2006.

[114] R. C. Petersen, R. Doody, A. Kurz, R. C. Mohs, J. C. Morris, P. V. Rabins, K Ritchie, M Rossor, L Thal, and B Winblad. Current concepts in mild cognitive impairment. *Archives of neurology*, 58(12):1985–1992, December 2001.

[115] Ronald C. Petersen and John C. Morris. Mild cognitive impairment as a clinical entity and treatment target. *Archives of neurology*, 62(7):1160–3; discussion 1167, July 2005.

[116] M. Philipose, K.P. Fishkin, M. Perkowitz, D.J. Patterson, D. Fox, H. Kautz, and D. Hahnel. Inferring Activities from Interactions with Objects. *IEEE Pervasive Computing*, 3(4):50–57, October 2004.

[117] D Podsiadlo and S Richardson. The timed "Up & Go": a test of basic functional mobility for frail elderly persons. *Journal of the American Geriatrics Society*, 39(2):142–148, 1991.

[118] Martha E. Pollack. Intelligent Technology for an Aging Population: The Use of AI to Assist Elders with Cognitive Impairment. *AI Magazine*, 26(2):9, June 2005.

[119] Martha E Pollack. The Use of AI to Assist Elders with Cognitive Impairment. *Association for the Advancement of Artificial Intelligence(AAAI)*, 26(2):9–24, 2005.

[120] Martha E Pollack. Intelligent Assistive Technology : The Present and the Future. In *User Modelling*, volume 4511, pages 5–6. Springer Berlin Heidelberg, 2007.

[121] Chase Randolph. *Repeatable Battery for the Assessment of Neuropsychological Status Update.* Psychological Corporation., San Antonio, Texas, 1998.

[122] Parisa Rashidi, Diane J Cook, Lawrence B Holder, and Maureen Schmitter-Edgecombe. Discovering Activities to Recognize and Track in a Smart Environment. *IEEE transactions on knowledge and data engineering*, 23(4):527–539, January 2011.

[123] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L. Littman. Activity recognition from accelerometer data. In *Proceedings of the 17th conference on Innovative applications of artificial intelligence*, pages 1541–1546. AAAI Press, July 2005.

[124] Nishkam Ravi, Dandeker Nikhil, Preetham Mysore, and Michael Littman. Activity recognition from accelerometer data. *In Proceedings of the Seventeenth Conference on Innovative Applications of Artificial Intelligence*, pages 1541—-1546, 2005.

[125] A. Reiss, D. Stricker, and G. Hendeby. Towards robust activity recognition for everyday life: Methods and evaluation, 2013.

[126] Saskia Robben, Mario Boot, Marije Kanis, and Ben Kr. Identifying and Visualizing Relevant Deviations in Longitudinal Sensor Patterns for Care Professionals. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 7th International Conference on*, pages 416–419, Venice, Italy, 2013.

[127] Saskia Robben, G. Englebienne, M.. Pol, and B Kröse. How Is Grandma Doing? Predicting Functional Health Status from Binary Ambient Sensor Data. In *2012 AAAI Fall Symposium Series*, pages 26–31, Washington D.C, 2012.

[128] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach.* Pearson Education, February 2003.

[129] Erik Scherder, Laura Eggermont, Dick Swaab, Marieke van Heuvelen, Yvo Kamsma, Mathieu de Greef, Ruud van Wijck, and Theo Mulder. Gait in ageing and associated dementias; its relationship with cognition. *Neuroscience and biobehavioral reviews*, 31(4):485–97, January 2007.

[130] Maureen Schmitter-Edgecombe, Courtney McAlister, and Alyssa Weakley. Naturalistic assessment of everyday functioning in individuals with mild cognitive impairment: the day-out task. *Neuropsychology*, 26(5):631–641, September 2012.

[131] Maureen Schmitter-Edgecombe, Carolyn Parsey, and Diane J. Cook. Cognitive correlates of functional performance in older adults: comparison of self-report, direct observation, and performance-based measures. *Journal of the International Neuropsychological Society JINS*, 17(5):853–864, 2011.

[132] Maureen Schmitter-Edgecombe, Carolyn Parsey, and Richard Lamb. Development and psychometric properties of the instrumental activities of daily living: compensation scale. *Archives of clinical neuropsychology : Journal of the National Academy of Neuropsychologists*, 29(8):776–92, December 2014.

[133] Maureen Schmitter-Edgecombe and Carolyn M Parsey. Assessment of functional change and cognitive correlates in the progression from healthy cognitive aging to dementia. *Neuropsychology*, 28(6):881–893, November 2014.

[134] Timothy Schmutte, Shelby Harris, Ross Levin, Richard Zweig, Mindy Katz, and Richard Lipton. The relation between cognitive functioning and self-reported sleep complaints in nondemented older adults: results from the Bronx aging study. *Behavioral sleep medicine*, 5(1):39–56, January 2007.

[135] Adriana Seelye, Nora Mattek, Diane Howieson, Thomas Riley, Katherine Wild, and Jeffrey Kaye. The Impact of Sleep on Neuropsychological Performance in Cognitively Intact Older Adults Using a Novel In-Home Sensor-Based Sleep Assessment Approach. *The Clinical neuropsychologist*, 29(1):1–14, February 2015.

[136] Megan G. Sherod, Randall H. Griffith, Jacquelynn Copeland, Katherine Belue, Sara Krzywanski, Edward Y. Zamrini, Lindy E. Harrell, David G. Clark, John C. Brockington, Richard E. Powers, and Daniel C. Marson. Neurocognitive predictors of financial capacity across the dementia spectrum: Normal aging, mild cognitive impairment, and Alzheimer's disease. *Journal of the International Neuropsychological Society : JINS*, 15(2):258–267, March 2009.

[137] Judith D Singer and John B Willett. *Applied Longitudinal Data Analysis.*, volume 102 of *Wiley Series in Probability and Statistics*. Oxford University Press, New York, 2003.

[138] Geetika Singla, Diane Cook, and Maureen Schmitter-Edgecombe. Recognizing independent and joint activities among multiple residents in smart environments. *Journal of Ambient Intelligence and Humanized Computing*, 1(1):57–63, March 2010.

[139] A. Smith. *Symbol Digit Modalities Test.* Western Psychological Services, Los Angeles, CA, 1991.

[140] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, August 2004.

[141] Peter J Snyder, Colleen E Jackson, Ronald C Petersen, Ara S Khachaturian, Jeffrey Kaye, Marilyn S Albert, and Sandra Weintraub. Assessment of cognition in mild cognitive impairment: a comparative study. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, 7(3):338–355, May 2011.

[142] W D Spector, S Katz, J B Murphy, and J P Fulton. The hierarchical relationship between activities of daily living and instrumental activities of daily living. *Journal of chronic diseases*, 40(6):481–489, January 1987.

[143] Erik Stone and Marjorie Skubic. Evaluation of an inexpensive depth camera for in-home gait assessment. *Journal of Ambient Intelligence and Smart Environments*, 3(4):349–361, December 2011.

[144] Stephanie Studenski, Subashan Perera, Dennis Wallace, Julie M Chandler, Pamela W Duncan, Earl Rooney, Michael Fox, and Jack M Guralnik. Physical performance measures in the clinical setting. *Journal of the American Geriatrics Society*, 51(3):314–22, March 2003.

[145] Yanmin Sun, Mohamed S. Kamel, Andrew K.C. Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378, 2007.

[146] Toshiro Suzuki and Sumio Murase. Influence of outdoor activity and indoor activity on cognition decline: use of an infrared sensor to measure activity. *Telemedicine journal and e-health : journal of the American Telemedicine Association*, 16(6):686–690, 2010.

[147] Cindy Woon Chi Tam, Linda Chiu Wa Lam, Helen F K Chiu, and Victor W C Lui. Characteristic profiles of instrumental activities of daily living in Chinese older persons with mild cognitive impairment. *American journal of Alzheimer's disease and other dementias*, 22(3):211–217, 2007.

[148] E L Teng and H C Chui. The Modified Mini-Mental State (3MS) examination. *The Journal of clinical psychiatry*, 48(8):314–8, August 1987.

[149] Holly Tuokko, Carolyn Morris, and Patricia Ebert. Mild cognitive impairment and everyday functioning in older adults. *Neurocase*, 11(1):40–47, February 2005.

[150] P. Turaga, R. Chellappa, V.S. Subrahmanian, and O. Udrea. Machine Recognition of Human Activities: A Survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, November 2008.

[151] Douglas L. Vail, Manuela M. Veloso, and John D. Lafferty. Conditional random fields for activity recognition. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems - AAMAS '07*, page 1, New York, New York, USA, May 2007. ACM Press.

[152] Joe Verghese, Richard B Lipton, Charles B Hall, Gail Kuslansky, Mindy J Katz, and Herman Buschke. Abnormality of gait as a predictor of non-Alzheimer's dementia. *The New England journal of medicine*, 347(22):1761–8, November 2002.

[153] Gilles Virone, Majd Alwan, Siddharth Dalal, Steven W Kell, Beverely Turner, John A Stankovic, and Robin Felder. Behavioral patterns of older-adults in assisted living. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 12(3):387–398, May 2008.

[154] Virginia G Wadley, Michael Crowe, Michael Marsiske, Sarah E Cook, Frederick W Unverzagt, Adrienne L Rosenberg, and Daniel Rexroth. Changes in everyday function in individuals with psychometrically defined mild cognitive impairment in the Advanced Cognitive Training for Independent and Vital Elderly Study. *Journal of the American Geriatrics Society*, 55(8):1192–1198, August 2007.

[155] L M Waite, D a Grayson, O Piguet, H Creasey, H P Bennett, and G a Broe. Gait slowing as a predictor of incident dementia: 6-year longitudinal data from the Sydney Older Persons Study. *Journal of the neurological sciences*, 229-230:89–93, March 2005.

[156] Shiaokai Wang, William Pentney, Ana-Maria Popescu, Tanzeem Choudhury, and Matthai Philipose. Common sense based joint training of human activity recognizers. pages 2237–2242, January 2007.

[157] Shuang Wang, Marjorie Skubic, and Yingnan Zhu. Activity density map visualization and dissimilarity comparison for eldercare monitoring. *IEEE Transactions on Information Technology in Biomedicine*, 16(4):607–614, July 2012.

[158] Jamie A. Ward, Paul Lukowicz, and Hans W. Gellersen. Performance metrics for activity recognition. *ACM Transactions on Intelligent Systems and Technology*, 2(1):1–23, January 2011.

[159] Sandra C Webber, Michelle M Porter, and Verena H Menec. Mobility in older adults: a comprehensive framework. *The Gerontologist*, 50(4):443–50, August 2010.

[160] D. Wechsler. *Adult Intelligence Test.* The Psychological Corporation, New York, USA, third edit edition, 2001.

[161] J.M. Williams. *Memory Assessment Scales professional manual.* Psychological Assessment Resources, Odessa, FL, 1991.

[162] R Williamson and B J Andrews. Gait event detection for FES using accelerometers and supervised machine learning. *IEEE transactions on rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society*, 8(3):312–319, September 2000.

[163] S. Willis and M. Marsiske. *Manual for the Everyday Problems Test.* Department of Human Development and Family Studies, Pennsylvania State University, 1993.

[164] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems).* Morgan Kaufmann Publishers Inc., June 2005.

[165] H J Woodford and J George. Cognitive assessment in the elderly: a review of clinical methods. *QJM : monthly journal of the Association of Physicians*, 100(8):469–484, August 2007.

[166] Ling Zhang, Beatriz C Abreu, Gary S Seale, Brent Masel, Charles H Christiansen, and Kenneth J Ottenbacher. A virtual reality environment for evaluation of a daily living skill in brain injury rehabilitation: reliability and validity. *Archives of physical medicine and rehabilitation*, 84(8):1118–1124, August 2003.

[167] A. L. Zulas, A. S. Crandall, M. Schmitter-Edgecombe, and D. J. Cook. Caregiver Needs from Elder Care Assistive Smart Homes: Nursing Assessment. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1):125–129, October 2012.