Smart Home-Based Prediction of Multidomain Symptoms Related to Alzheimer's Disease

Ane Alberdi , Alyssa Weakley, Maureen Schmitter-Edgecombe, Diane J. Cook, *Fellow, IEEE*, Asier Aztiria, Adrian Basarab , and Maitane Barrenechea

***Abstract*—As members of an increasingly aging society, one of our major priorities is to develop tools to detect the earliest stage of age-related disorders such as Alzheimer's Disease (AD). The goal of this paper is to evaluate the possibility of using unobtrusively collected activity-aware smart home behavior data to detect the multimodal symptoms that are often found to be impaired in AD. After gathering longitudinal smart home data for 29 older adults over an average duration of >2 years, we automatically labeled the data with corresponding activity classes and extracted time-series statistics containing ten behavioral features. Mobility, cognition, and mood were evaluated every six months. Using these data, we created regression models to predict symptoms as measured by the tests and a feature selection analysis was performed. Classification models were built to detect reliable absolute changes in the scores predicting symptoms and SmoteBOOST and wRACOG algorithms were used to overcome class imbalance where needed. Results show that all mobility, cognition, and depression symptoms can be predicted from activity-aware smart home data. Similarly, these data can be effectively used to predict reliable changes in mobility and memory skills. Results also suggest that not all behavioral features contribute equally to the prediction of every symptom. Future work therefore can improve model sensitivity by including additional longitudinal data and by further improving strategies to extract relevant features and address class imbalance. The results presented herein contribute toward the development of an early change detection system based on smart home technology.**

***Index Terms*—Activity recognition, Alzheimer's disease, automatic assessment, behavior, multimodal symptoms, older adults, smart homes.**

Manuscript received July 19, 2017; revised October 25, 2017 and December 24, 2017; accepted January 21, 2018. Date of publication January 25, 2018; date of current version October 15, 2018. This work was supported by the National Institutes of Health under Award R01EB015853. *(Corresponding author: Ane Alberdi Aramendi.)*

A. Alberdi, A. Aztiria, and M. Barrenechea are with the Department of Electronics and Computing, Arrasate 20500, Spain (e-mail: aalberdiar@mondragon.edu; aaztiria@mondragon.edu; mbarrenetxea@mondragon.edu).

A. Weakley and M. Schmitter-Edgecombe are with the Department of Psychology, Washington State University, Pullman, WA 99164 USA (e-mail: alymae@wsu.edu; schmitter-e@wsu.edu).

D. J. Cook is with the School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164 USA (e-mail: djcook@wsu.edu).

A. Basarab is with the Université de Toulouse, Institut de Recherche en Informatique de Toulouse, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5505, Université Paul Sabatier, 31062 Toulouse, France (e-mail: adrian.basarab@irit.fr).

Digital Object Identifier 10.1109/JBHI.2018.2798062

I. Introduction

INCREASING life expectancy in developed countries has resulted in a growing number of cases of people affected by age-related neurodegenerative diseases, such as Alzheimer's Disease (AD). An estimate of 115.4 million people will suffer from AD in 2050 [1], which can result in devastating consequences in terms of health-care costs and quality of life of patients and caregivers. While there is no known cure [2], treatments to delay and reduce cognitive and behavioral symptoms of AD do exist and are demonstrated to be more effective the sooner they are applied [3]. Therefore, as a matter of general interest, the search for methods of early detection is currently a high priority issue. Such methods could lead to earlier detection and therefore more effective intervention. The resulting benefits include an increase in the independence of the patients, an improvement in quality of life for them and their caregivers and a reduction in health-care costs.

Although AD's clinical hallmark is episodic memory impairment [4], it manifests symptoms in multiple domains, including mood, behavior, and cognition [5]. These symptoms and the associated pathology are usually measured by means of self- and informant- report questionnaires, clinical assessments conducted by health care professionals and medical examinations that may involve brain imaging. Often evaluations are initiated after symptoms have been prominent for some time, resulting in a delayed diagnosis [6]. Given that AD pathology in the brain accumulates slowly over time, a key for the treatments to be effective is early detection of the disease and implementation of available treatments.

Smart homes are an emerging technological solution, based on the use of embedded sensors to enhance homes' intelligence, enabling the unobtrusive monitoring of resident's behavior [7]. Real-life data can be gathered non-stop in a completely naturalistic way, offering a complete and ecologically valid view of older adults' behavior and allowing the detection of changes that might indicate the onset of a disorder. If smart home-based behavior shifts were mapped to AD, many disadvantages of the usual assessment methods could be overcome: detection could be made without the need for older adults to travel to a health center to receive expensive and invasive diagnostic testing. In contrast, smart home monitoring may detect cognitive changes as they occur, resulting in less expensive and more timely diagnosis.

In order to map detected behavior shifts to AD symptoms, machine learning-based models can be used. Machine learning is a

TABLE I
OVERVIEW OF RELATED WORK CATEGORIZED IN TERMS OF THE MEASURED BEHAVIORAL FEATURES, THE EMPLOYED ASSESSMENT TESTS, PREPROCESSING AND ANALYSIS TECHNIQUES, AND THE OBSERVED RESULTS (C = COGNITION, MOB = MOBILITY, M = MOOD, n = SAMPLE SIZE)

Ref.	Behavior	Tests	Preprocessing and Analysis	Main results
Dawadi *et al.* ($n = 18$)	ADL, sleep, mobility, outings	C: RBANS; Mob: TUG	-AR + daily behavior stats-Machine-learning	C: r = 0.72, %72 Mob: r = 0.45, %76
Hayes *et al.* ($n = 14$)	Amount of activity, walking speed	C: CDR, MMSE	-Wavelet analysis -Mixed-model ANOVA	MCI Doubled coefficient of variation in the median walking speed ($p < 0.03$) and increased variability in the amount of activity ($p < 0.008$)
Galambos *et al.* ($n = 5$)	Time out, activity level	C: MMSE, SFHS-12; M: GDS	Motion and out of home density maps	Correlation between the scores and activity level/outings (Qualitative)
Petersen *et al.* ($n = 85$)	Time out, in-home walking speed	C: CDR; M, Physical activity	Tobit mixed-effects regression model	Correlated time spent out of home and cognitive ($p < 0.001$), physical ($p < 0.001$) and emotional state ($p < 0.001$).
Austin *et al.* ($n = 16$)	Time out, n° of phone calls, computer use, walking speed, mobility	M: loneliness	Longitudinal linear-mixed effects regression model + CV	Correlated loneliness and both time out of home ($p < 0.01$) and computer sessions ($p < 0.05$)
Alberdi *et al.* ($n = 29$)	ADL, sleep, mobility, global routine, outings	C: RBANS, PRMQ, Digit Cancel; Mob: Arm Curl, TUG; M: GDS	-AR + daily behavior stats + RCI + positive/negative change -Machine-learning + SMOTEBoost + wRACOG	See Results in Section III

subdiscipline of artificial intelligence (AI) aimed at building algorithms that are able to learn and/or adapt their structure based on a set of observed data (i.e., example data or past experience) [8], [9]. This technique offers an approach for the analysis of high-dimensional and multimodal biomedical data. A wide variety of methods exist within this area, including both regression (e.g. Support Vector Regression, Linear Regression or k Nearest Neighbors) and classification methods (e.g., Support Vector Machines, AdaBoost, Multilayer Perceptron or Random Forest). Whereas regression models predict continuous variables (e.g., a score for a standardized assessment test), classification models determine symbolic class labels for the data (e.g., affected vs. non-affected by a disease). For a detailed explanation of specific machine learning algorithms, we refer the reader to the literature [10], [11].

Our goal in this paper is to assess the possibility of detecting changes in psychological, cognitive and behavioral symptoms of AD by making use of unobtrusively collected smart home behavior data and machine learning techniques. The affirmation of this hypothesis would result in development and implementation of an early detection system for disorders that provoke behavioral changes, such as AD. Such a system could alert patients and relatives of likely changes, making it possible to take timely action.

Previous research has demonstrated that the combination of machine learning techniques and longitudinal monitoring of smart home-based behavioral data can be useful not only to assess older adults' health states but also to detect onset and monitor progression of some age-related diseases and disorders. Dawadi *et al.* found that the overall cognitive and mobility states of older adults could be predicted from unobtrusively collected in-home behavior data [12]. For that purpose, they introduced an algorithm called Clinical Assessment using Activity Behavior (CAAB) and tested its validity for global cognition (measured by the Repeatable Battery for the Assessment of Neuropsychological Status, or RBANS) and mobility (measured by the Timed Up and Go, or TUG) using time series-based descriptive statistics of daily activities. Hayes *et al.* [13] found Mild Cognitive Impairment (MCI), as measured by the Clinical Dementia Rating (CDR) and Mini-Mental State Examination (MMSE) tests, to be correlated with in-home walking parameters and mobility measures. MCI implies cognitive decline in one or more domains of cognition (e.g., memory, language, executive function) that is greater than what could be attributed to normal aging, but does not meet the threshold for a diagnosis of a dementia disorder like AD [14].

In related work, Galambos *et al.* [15] discovered associations between overall in-home activity and outing patterns with both dementia and depression, which is also known to be a common AD symptom. The Geriatric Depression Scale (GDS), as well as the MMSE and Short Form Health Survey-12 scales were used to determine subjects' state. Petersen *et al.* [16] also found emotional states, specifically mood and loneliness, to be correlated to outing patterns, whereas they also verified the possibility of predicting other overall health predictors such as physical activity from these data. Austin *et al.* also predicted the loneliness of older adults by analyzing behavioral data [17]. A comparative summary of the sample sizes, techniques used, symptoms predicted, and observed results are given in Table I.

Nonetheless, there's still much work to do towards the development of models to reliably detect AD symptoms from unobtrusively collected in-home behavioral data. The predictability of the wide range of multi-modal symptoms of AD is yet to be analyzed, as well as the contribution of many behavioral traits to these models. Moreover, the possibility of detecting a Reliable Change [18] in AD multimodal symptoms from smart home data is yet to be researched. In addition, solutions have not been heavily explored to handle imbalanced class distributions (i.e., a much larger number of negative cases than positive cases) that are common in such environments. Furthermore, there are few studies where quantitative detection results have been given.

This paper aims at filling this research gap. Previous work has demonstrated the validity of daily behavioral statistics for the prediction of cognitive and mobility skills of older adults [12]. Building on this foundation, we will introduce new behavioral features and will analyze their validity for the detection of reliable changes in multi-modal AD symptoms.

The main contributions of this work can be summarized as follows. We analyze the predictability of several multi-domain symptoms often found to be impaired in AD, we analyze the contribution of behavioral features to the prediction of these health assessment scores, and we introduce and assess new smart home-based behavior features to quantify global daily routine. In addition, we offer an approach to detect a reliable change in health assessment scores based on unobtrusively collected behavioral data and to address the accompanying imbalanced class distribution problem.

II. Methods

A. Data Collection

First, we unobtrusively collected in-home behavioral data for older adults living in smart homes in two senior-living communities and we gathered corresponding biannual neuropsychological assessment data. This data was collected by the Center for Studies in Adaptive Systems (CASAS) and the Neuropsychology and Aging Laboratory at Washington State University (WA, USA). Review and approval by the Washington State University Institutional Review Board was obtained for the study. Part of this data ($n = 18$ older adults) was analyzed in previous work [19]. For this work, a larger sample is available thanks to a longer monitoring time and to the inclusion of more subjects in the study.

The current study focuses on cognition, mobility, and mood (depression) scores (see Table III), which were collected as part of the biannual assessment and have been found to be affected by AD [5]. Cognitive abilities of the older adults were measured by means of the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) [20], the Prospective and Retrospective Memory Questionnaire (PRMQ) [21] and a Digit Cancellation test, while mobility was assessed by Timed Up and Go (TUG) [22] and Arm Curl [23] tests. Whereas the RBANS is a brief, individually administered battery to measure cognitive decline or improvement across several domains (Immediate Memory, Visuospatial, Language, Attention and Delayed Memory), PRMQ is a 16-item self-report measure of prospective and retrospective memory slips in daily life. The Digit Cancellation test is a user-friendly assessment of various aspects of prefrontal cortex functioning (namely, information processing speed, the ability to focus attention and executive functioning) [24]. TUG and Arm Curl are physical tests to measure patients' risk for falling and upper body strength, respectively. The Geriatric Depression Scale - Short Form (GDS-15) [25] was used to assess the depression level of the participants. The 15-item GDS is the reduced version of the original 30-item GDS scale, which is a screening measure used to detect clinical levels of depression in older adults. A score of 10 or greater is suggestive of clinical depression.

TABLE II
Participants' Characteristics

Cognitive status	Healthy	At risk	Difficulties
Group Size	$N = 13$	$N = 10$	$N = 6$
Age	82.85 (73-92)	86.20 (73-97)	84.50 (82-90)
Education	17.58 (16-20)	17.20 (12-20)	17.67 (16-20)
Gender	$m = 4, f = 9$	$m = 3, f = 10$	$m = 1, f = 5$

($m =$ male, $f =$ Female. Age and Education are Specified by Mean (Range).)

TABLE III
Modality, Test-Retest Reliability, and Standard Deviations of the Scores Used in the Study

Domain	Score	r_{score}	SD_{score}	Ref.
Mobility	Arm Curl	0.96	4.98	[27]
	TUG	0.96	3.18	[26]
	Digit Cancellation	0.85	37.20	[28]
	RBANS:			
	+total	0.84	15.58	[29]
	+attention	0.16	19.00	
	+delayed memory	0.77	13.29	
	+immediate memory	0.75	16.58	
Cognition / Memory	+visuospatial	0.76	15.31	
	+language	0.33	15.31	
	PRMQ:			
	+total	0.89	9.15	[30]
	+prospective memory	0.85	4.91	
	+retrospective memory	0.89	4.98	
Mood	GDS	0.68	2.20	[31]

The smart home sensor data used for this study was collected from 2011 through 2016, a period in which the data were collected continuously for durations ranging from <1 month to 60 months ($M = 19.95$ months, SD $= 17.98$ months) depending on the residence. Health assessment data was also collected for 29 of the older adults who were living independently in the smart homes. Participants were classified as either cognitively healthy, at risk for cognitive difficulties or experiencing cognitive difficulties. See Table II for group demographic information. Participants in the cognitive risk group had lowered performance on one or more cognitive tests (relative to an estimate of premorbid abilities), but did not meet criteria for MCI or dementia. One participant in the cognitive difficulties group was diagnosed with a brain tumor with marked reductions in cognition following diagnosis. The remaining participants in the cognitive difficulties group met criteria for mild cognitive impairment (MCI) as outlined by the National Institute on Aging-Alzheimer's Association workgroup [26].

B. Preprocessing

1) Day-Level Behavior Feature Extraction: Smart homes were set up to collect all sensor events that took place in each residence during the study period. Each data stream entry described a single sensor event in terms of the event's timestamp, ID of the sensor detecting the event, and type of event (activation/deactivation).

Note that, a raw-sensor data entry by itself is meaningless: the same sensor event can occur when performing different

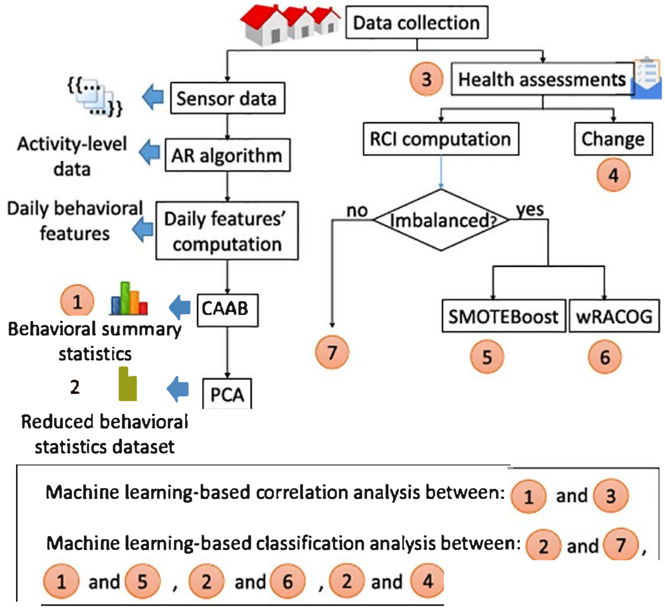
Data collection
{{...
...}}
Sensor data
3
Health assessments
Activity-level data
AR algorithm
RCI computation
Change
4
Daily behavioral features
Daily features' computation
no
Imbalanced?
yes
1
Behavioral summary statistics
CAAB
SMOTEBoost
wRACOG
7
5
6
2
PCA
Reduced behavioral statistics dataset
Machine learning-based correlation analysis between: 1 and 3
Machine learning-based classification analysis between: 2 and 7 ,
1 and 5 , 2 and 6 , 2 and 4

Fig. 1. Overview of the research methods.

activities and multiple occurrences of a specific activity may yield different event sequences. Therefore, in order to interpret the event data, it was first necessary to assign an activity label to each sensor entry, taking into account the context in which the sensor event occurred. For that purpose, the AR Activity Recognition algorithm [32] was used. This algorithm maps each of the sensor events to a value from a predefined set of activity labels in real-time, by applying an adaptive-length sliding window to the raw sensor data stream. The predefined set of activities include both ambulatory activities (such as mobility inside the home) and specific activities of daily living (ADLs), which were encoded by numbers from 1 to 12 (i.e., Sleep = 1, Cook = 2, Relax = 3, ..., Other = 12). This approach not only takes into account recent sensor events but also contextual information such as the activity label that was assigned to the previous time window. The reliability of this algorithm has been demonstrated in previous work, where accuracy greater than 98% was achieved on 30 testbed smart homes using three-fold cross validation [32]. Fig. 2 shows an extract of an AR activity labeled sensor data stream.

Once activity-level information was available, we computed 17 daily behavior features for each subject, explaining their daily sleep and mobility patterns, time spent in several specific ADLs (e.g., cook, eat) and overall characteristics of their routines. A detailed list of the computed features can be seen in Table IV.

```
2011-06-18 13:23:16.33 WorkArea WorkArea M005 ON Work
2011-06-18 13:23:18.04 WorkArea WorkArea M005 OFF Relax
2011-06-18 13:23:18.58 WorkArea WorkArea M005 ON Relax
2011-06-18 13:23:24.95 WorkArea WorkArea M005 OFF Relax
2011-06-18 13:23:31.53 Kitchen Kitchen MA006 ON Cook
2011-06-18 13:23:34.53 Kitchen Kitchen MA006 OFF Cook
2011-06-18 13:23:35.46 Kitchen Kitchen MA006 ON Cook
2011-06-18 13:23:37.72 Kitchen Kitchen MA006 OFF Cook
2011-06-18 13:23:55.33 Kitchen Kitchen MA006 ON Cook
2011-06-18 13:23:56.45 Kitchen Kitchen MA006 OFF Cook
2011-06-18 13:24:03.53 Kitchen Kitchen MA006 ON Cook
2011-06-18 13:24:05.26 Kitchen Kitchen MA006 OFF Cook
2011-06-18 13:24:11.08 WorkArea WorkArea M005 ON Eat
2011-06-18 13:24:18.59 WorkArea WorkArea M005 OFF Eat

2011-07-28 08:40:39.41 Bedroom Bedroom MA007 ON Sleep
2011-07-28 08:40:41.82 Bedroom Bedroom MA007 OFF Sleep

2011-07-29 12:22:06.83 WorkArea WorkArea M005 ON Work
2011-07-29 12:22:08.69 WorkArea WorkArea M005 OFF Work
```

Fig. 2. Extract of an AR activity-labeled sensor event data stream.

TABLE IV
DAY-LEVEL ACTIVITY FEATURES INCLUDED IN THE STUDY

Type	Day-level features
Duration of specific activities (6 features)	Time spent per day in cooking, eating, relaxing, and performingpersonal hygiene and nighttime toileting activities as well as time out of the home.
Sleep-related (2 features)	Daily sleep duration and frequency.
Mobility-related (2 features)	Total number of activated sensors and total distance covered walking inside the home per day.
Routine-related (7 features)	Complexity of the daily routine, number of total and of non-repeated activities performed per day, maximum and minimum inactivity times, day length and similarity with the previous day.

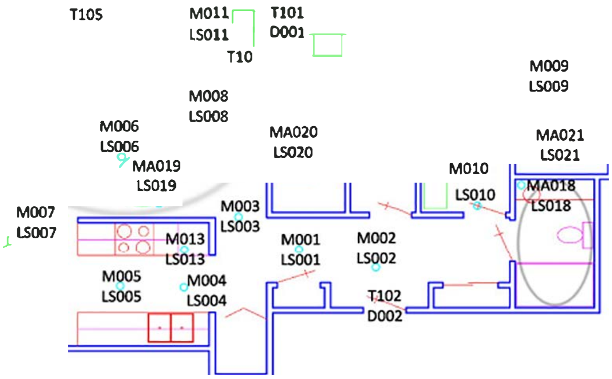

Fig. 3. Floor plan and sensor layout of one of the smart home sites.

The daily distance that the subjects traveled inside their homes was estimated by computing the distance between areas of the home covered by each passive infrared (PIR) motion sensor as determined from based on the floor plan and sensor layout (see example in Fig. 3). Three of the apartments lacked specific information about the positioning of the sensors within the houses. In those cases, it was first necessary to estimate the positions of the sensors, which was done by considering these apartments to be of a similar shape to the rest and checking the activation order of the sensors in the raw sensor data files. Once all sensor positioning information was available, we computed the daily sum of the Euclidean distances between the consecutively-activated motion sensors in order to estimate the total walking distance traveled by the inhabitants. Note that this approach only provides an approximation of the real covered distance, as it does not consider the existence of walls or other obstacles between the sensors that must be avoided or navigated.

To compute daily-routine features, we first extracted the daily activity sequence from the AR-labeled sensor data stream. Shannon entropy was used to measure the complexity of the daily routine. To compute this entropy value, we estimated the daily probability distribution (histogram) of the activity sequence (P) and then applied the entropy formula shown in Equation 1,

$$Complexity_{routine} = \sum_{activity=1}^{12} P_{activity} - \log_2 P_{activity} \tag{1}$$

where $P_{activity}$ was the probability of a certain activity to occur for a given day based on the actual day's histogram.

The same encoded activity sequence was used to compare the daily routines of consecutive days. For this purpose, we used an implementation of the "gestalt pattern matching" algorithm [33]. This SequenceMatcher function, available in Python, expresses the similarity of any two sequences as a value between 0 and 1. We use this function to determine the degree of similarity between consecutive days. Finally, we checked the timestamps of the daily activity events and computed the day-length as the time elapsed between the first and the last detected activity of the day. The remaining features in Table IV are self-explanatory.

2) Between-Assessments Behavior Statistics' Computation: The previous step yielded a set of daily activity features for each subject. We then applied the CAAB algorithm, which was introduced earlier [19], using RStudio for R to the daily activity data in order to extract behavior statistics for each between-assessment period.

In summary, the CAAB algorithm was used to apply the following processing steps to the daily behavior data: 1) Take each subject's between-assessment daily behavior data (which was 6 months in length as assessments were performed twice a year), 2) Apply a log transform and a Gaussian detrending to each time-series (behavioral variable), 3) Compute five summarizing time-series statistics (variance, skewness, kurtosis, autocorrelation, and change) for each behavioral feature in this period using a sliding window of length 7 days, and 4) Compute the 6-month average of each time-series statistic and use the set of averages for the final predictions.

The resulting preprocessed dataset for further analysis was a collection of 85 (5 time-series statistics of 17 behavioral features) biannual summary behavior statistics for each of 29 older adults who were living alone in their sensorized apartments for a period of $24.0 \pm 13.68(SD)$ months.

3) Health Assessment Scores: Our goal is to create prediction models that map smart home-based behavior features to health assessment values that might capture AD symptoms. In this study, our target variables are the Arm Curl and TUG mobility test scores, cognition assessment based on Digit-Cancellation test, RBANS and PRMQ scores and subscores, as well as depression symptoms represented as GDS test-scores. All these values were collected from the participants at the end of each corresponding 6-month period.

Self-reported scores are usually strongly subject dependent. In addition, two people might achieve different results in the same test even if they have similar skills, due to their intrinsic characteristics. As a measure to avoid this inter-subject variability in the scores, we used a standardization method based on the Reliable Change Index (RCI) [18] computation. RCI compares assessment scores for each participant at one time point to previous scores for the participant to determine whether the participant has undergone a significant change in his/her performance. Detecting a significant change implies that the subject's scores have changed sufficiently (exceeding a specified threshold) so that the change is unlikely to be due to measurement unreliability (i.e., due to repeat testing or practice effects). We looked for two types of reliable absolute changes: the first one compares each assessment value to the participant's baseline values ($RCI_{baseline}$), whereas the second one compares each assessment point to the same participant's previous assessment point ($RCI_{consecutive}$).

In order to calculate the RCIs for the scores used herein, we gathered test-retest reliability (r_{score}) and standard deviation (SD_{score}) values that the tests have shown in their development cohorts and/or in previous work, as shown in Table III. Therefore, the RCIs for each subject were computed as:

$$RCI_{baseline}(i) = \frac{Score_i - Score_{baseline}}{\sqrt{2SEm}} \tag{2}$$

$$RCI_{consecutive}(i) = \frac{Score_i - Score_{i-1}}{\sqrt{2SEm}} \tag{3}$$

where SEm or Standard Error of Measurement represents the expected variation of the observed test scores due to measurement error and is computed as $SEm = SD_{score}\sqrt{1 - r_{score}}$, r_{score} is the test-retest reliability measuring the consistency of the test scores over time, $Score_i$ is the test score at assessment point i, $Score_{baseline}$ is the test score at the first/baseline assessment and $Score_{i-1}$ is the test score at the previous assessment point.

Some of the assessment scores result in very few positive instances (data instances where a reliable change was observed), resulting in highly imbalanced class data. For the following analyses, we removed from the study those tests which were extremely imbalanced (<5% of positive instances). We distinguished the remaining tests as imbalanced (5%–30% of positive instances) and balanced data (30%–50% of positive instances).

Additionally, we also considered the possibility of detecting improvement and decline in test scores among consecutive assessment points as a method to reduce inter-subject variability. Comparing an individual's score to his/her own previous one allows us to standardize the results, since it is a way to evaluate the improvement or decline of each individual's skills in the time period under analysis, regardless of the absolute values of the scores. In this case, the difference between each consecutive assessment point was computed for each self-reported test score of each subject. Every data instance with an improvement in the scores ($\geqslant$0) was considered as a positive point whereas a decline in the performance of the skill being evaluated by tests (<0) was labeled as a negative point.

TABLE V
TASK-SPECIFIC GROUPING OF DAILY FEATURES

Group	Day-level features
Daily-routine	Complexity of the daily routine, number of total activities and number of non-repeated activities performed per day, maximum and minimum inactivity times, day length and similarity with the previous day
Mobility	The total number of activated sensors and the total distance covered walking inside the apartment per day
Outings	Time spent per day in being out of home
Mobility & outings	Mobility + Outings
Sleep	The daily sleep duration and frequency
Overnight toileting	Time spent per day in nighttime toileting activities
Overnight patterns	Sleep + Overnight toileting
Cook & eat	Time spent per day in cooking and eating

C. Cognition and Mobility Change Prediction

The preprocessed dataset was analyzed using Weka [34], a free machine learning software written in Java. First, we performed a correlation analysis between the mobility, cognition, and mood assessment scores and the smart home behavior data. For this purpose, we used four different regression models using all behavior features computed in the previous step for each one of the scores. The four models we evaluated were Support Vector Regression (SVr) with a linear kernel, Linear Regression (LinearR), SVr with a Radial Basis Function (RBF) kernel and k nearest neighbors (kNN) algorithms. We compared the correlation coefficients (r) and Mean Absolute Errors (MAE) of the models using 10-fold cross validation (CV) approach. Corresponding pairwise random algorithms were built and evaluated in our dataset following the same process. These random algorithms provided a basis of comparison to ensure that performance results are not due to chance. The random algorithms were built using a uniformly distributed random data-matrix of the same size as the real behavioral data while respecting each variable's data range as in the original dataset. A corrected paired t-test was used to detect a significant improvement of smart home-based algorithms in comparison to the random data algorithms. Adjusted p-values ($^{*}p < 0.01$, $^{**}p < 0.001$) were used to avoid Type 1 error when checking for significance.

In order to analyze the types of behavior features that are most correlated with each one of the tests, we built activity-specific models for the main test scores with a linear SVr and evaluated the models using 10-fold cross validation. The behavior features that were included in each one of the models are shown in Table V. Again, we searched for statistically significant improvement in comparison to pairwise random algorithms using a corrected paired t-test and adjusted p-values ($^{*}p < 0.01$, $^{**}p < 0.001$).

Regarding RCI detection, we used different approaches for the imbalanced and balanced datasets. First, balanced datasets containing all behavioral features were reduced by means of a Principal Component Analysis (PCA). PCA is a popular statistical technique based on the projection of the data to a lower-dimensional subspace, useful for finding patterns in high-dimensional datasets [35]. Principal Components that explained 95% of the variability in the behavior data were kept to create the reduced datasets. The SVM, AdaBoost, Multilayer Perceptron (MLP) and Random Forest (RF) algorithms were trained and validated using ten-fold cross validation. Evaluation metrics include area under the ROC curve (ROC_{auc}), area under the Precision-Recall curve (PRC_{auc}), $Fscore$, and sensitivity. The combination of these metrics offers an excellent overview of both the models' overall performance and the capability to detect the event of interest (the reliable change event), and are especially suitable when the data distribution is skewed. A corrected paired t-test was used to detect a significant improvement of smart home-based algorithms in comparison to the pairwise random data algorithms, and an adjusted p-value ($^{*}p < 0.0125$) was employed to avoid Type 1 error.

For the imbalanced datasets, a different approach was required. Common machine-learning algorithms tend to create models that are biased towards the majority class when being applied to imbalanced datasets, resulting in high accuracies but very low sensitivity. In most of the health-related machine learning applications, the events in which we are more interested are the rare events or the minority class, highlighting the need to use alternative methods to improve the detection of these minority events. Two algorithmic approaches are tested in the current work to overcome this issue. The first one, SMOTEBoost [36], is a method combining boosting techniques with SMOTE [37] oversampling techniques. Whereas boosting aims at creating a "strong" classifier using a set of "weak" classifiers, SMOTE is a technique that oversample the minority class by creating synthetic data instances and thus reducing class imbalance. SMOTEBoost combines these processes iteratively in order to improve the sensitivity of the models without affecting the overall accuracy.

The second approach, the wrapper-based Rapidly Converging Gibbs sampler (wRACOG) [38], is a minority-class oversampling algorithm based on Gibbs sampling. Unlike SmoteBOOST and most of the minority-class oversampling techniques, wRACOG takes into account the underlying probability distribution of the minority class and the interdependencies of the data attributes when synthetically generating rare-event samples. This results in a better representation of the minority class. Moreover, wRACOG learns the models iteratively, selecting from the Markov chain generated by the Gibbs sampler the samples that have the highest probability of being misclassified by a learning model (wrapper) at each step, often leading to better classification rates. wRACOG stops iterating when there is no further improvement with respect to a chosen performance metric.

First, we built prediction models for imbalanced datasets using SMOTEBoost and kNN with $k = 5$ as the "weak" classifier which we validated using 3-fold cross validation. Pairwise random algorithms were also built using the previously-mentioned random data and were validated for prediction of our data following the same 3-fold CV process. Again, we computed ROC_{auc}, PRC_{auc}, $Fscore$ and sensitivity of the models. McNemar's test was applied to check whether a significant

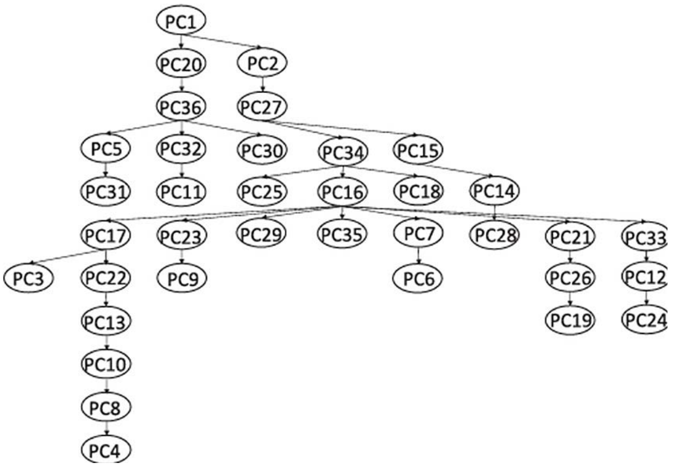

Fig. 4. Chow-Liu tree for the PCA-reduced dataset.

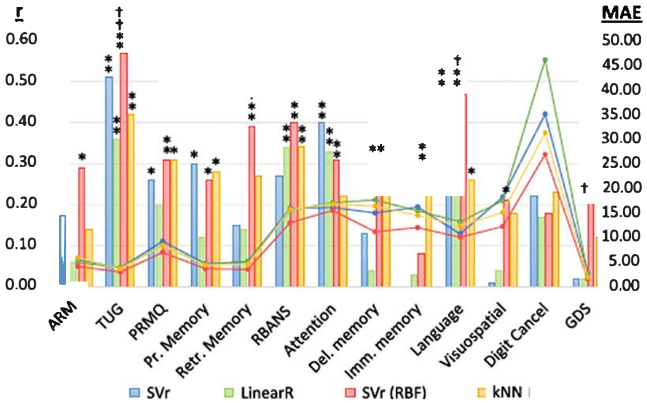

Fig. 5. Regression results for the absolute test scores using all behavioral features based on 10-fold CV (statistically significant improvement for r (adjusted $^{*}p < 0.01$, $^{**}p < 0.001$) and for MAE ($^{\dagger}p < 0.01$, $^{\dagger\dagger}p < 0.001$) in comparison to the corresponding pairwise random algorithm)). Bars represent r and lines represent MAE.

improvement (for an adjusted p-value ($^{*}p < 0.005$)) was observed using smart home-based prediction of reliable change in the scores in comparison to random data algorithms.

Next, we built the prediction models for the same imbalanced datasets following the second approach, i.e., using the wRACOG algorithm. For this purpose, it was first necessary to discover the interdependencies of the data attributes. In order to reduce the dimensionality of the data and to make it easier to map the interdependencies between the attributes, we used the PCA-based reduced datasets explaining the 95% of the data variance. Moreover, wRACOG assumes that the data attribute values are categorical, so we first discretized all of the principal components (PCs) into five uniform bins. We then constructed the Bayesian tree of dependencies following the Chow-Liu algorithm in Weka. The Chow-Liu algorithm [39] aims at constructing a maximal weighted spanning tree in a graph, allowing each attribute to have exactly one parent on which its value depends. Thus, the interdependencies between the PCs were discovered. Fig. 4 shows the Chow-Liu interdependency tree for the PCA-reduced and discretized baseline dataset.

A kNN algorithm was used as the wrapper classifier and two different stopping criteria for the iterative process were tested: 1) First, as in many applications where the detection of the reliable change might be critical, we searched for the maximum sensitivity of the models. 2) Second, for cases where the overall prediction ability of the models might be more interesting, we used the maximized ROC_{auc} metric as the stopping criteria for the algorithm. A 5-fold CV was performed for validation purposes and ROC_{auc}, PRC_{auc}, $F score$, and sensitivity of the models were evaluated. As in previous cases, in order to check for statistically significant smart home-based prediction of reliable change in the scores, we compared model outputs to those of their pairwise random algorithms by means of a McNemar's test. An adjusted p-value ($^{*}p < 0.005$) was used to avoid family-wise (Type 1) error rate. The PCA-reduced random dataset was discretized following the same process as the actual smart home dataset.

Finally, for the detection of a person's improvement/decline from smart home data, we used the PCA-based reduced dataset as in the previous case. The SVM, AdaBoost and RF algorithms were trained and validated following a 10-fold CV approach to discriminate the positive class (a score improvement)) from the negative class (a score decline). ROC_{auc}, PRC_{auc}, and $F score$ were computed for each one of the algorithms and compared to the values of their pairwise-random algorithms. As the detection of a decline in self-reported skill performance might be more important than the detection of an improvement, we also computed the sensitivity of the algorithms towards these negative events. All statistical significances were checked for adjusted p-values ($^{*}p < 0.01$, $^{**}p < 0.001$). Fig. 1 provides an overview of the research methods.

III. Results

A. Absolute Test Scores' Prediction

Fig. 5 shows the results of predicting the absolute test scores using all smart home behavioral features with regression learners. For mobility tests, whereas Arm Curl had low correlation with behavioral data, TUG demonstrated a moderate to strong correlation. For the cognition overall scores and subscores, the measures showed mostly moderate correlations with behavioral data. Exceptions included the visuospatial and immediate memory subscores of the RBANS test and the digit cancellation test scores, which were found to correlate weakly. In fact, the digit cancellation test didn't show any statistically significant improvement compared to random models, whose MAE is also the highest. Finally, depression showed a weak correlation with the global set of smart home behavioral data.

Regressions based on specific activities, which can be seen in Fig. 6, showed some interesting results. The Arm Curl mobility test showed weak but statistically significant correlations with outings, and cooking and eating features. In contrast, the TUG test showed significant moderate correlations with daily routines, overnight toileting and the combination of overnight toileting and sleep, as well as a significant weak correlation with cooking and eating features.

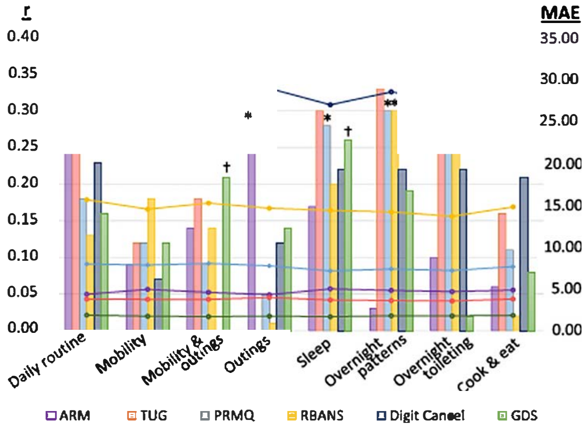

Fig. 6. Regression results for the absolute test scores by behavior feature type based on 10-fold CV (statistically significant improvement for r (adjusted $^{*}p < 0.01$, $^{**}p < 0.001$) and for MAE ($^{\dagger}p < 0.01$) in comparison to the corresponding pairwise random algorithm)). Bars represent r and lines represent MAE.

Regarding the self-report questionnaire, the global PRMQ score was moderately associated with daily routine and with the overnight patterns, as well as weakly correlated with sleep and overnight toileting. RBANS was moderately correlated with overnight patterns, whereas it was also showing weak yet statistically significant correlations with mobility, daily routine, and overnight toileting behaviors. Digit Cancellation processing speed was found to be moderately correlated with sleep and overnight patterns, and weakly yet significantly correlated with overnight toileting features.

Finally, for the geriatric depression assessment, we did not find any significant correlations but we perceived a significant reduction of the MAE of the models for mobility alone as well as for the mobility, outings, and sleep feature sets.

The overview of the trends shows that sleep and overnight behavior patterns, together with daily routine features presented in this paper, are the behavioral features that contribute the most to the prediction of the several health assessments.

B. RCI Detection

The detection of reliable change on attention and language skills were excluded from our objectives due to the uncertainty that their low test-retest reliability would introduce in the results obtained for these labels. Global PRMQ and subscores, consecutive global RBANS scores, RBANS subscores related to immediate memory, Digit cancellation, and the GDS test score were excluded from the RCI detection analyses as they were capturing less than 5% of the reliable change instances. Among the remaining labels, only the reliable change in Arm Curl scores from baseline had enough positive instances to be considered a balanced dataset. The remaining scores (RBANS, RBANS delayed memory, RBANS visuospatial and TUG change from baseline, and RBANS delayed memory, RBANS visuospatial and TUG change between consecutive assessments) were considered imbalanced and were processed as such.

Table VI shows the results for Arm Curl reliable change detection from baseline using 37 PCs explaining the 95% variability of the data. All four classifiers demonstrated a statistically significant improvement in terms of Area under the PR curve, $Fscore$ and sensitivity for the adjusted p-value, whereas area under the ROC curve showed reasonable results surpassing the 0.6 barrier.

TABLE VI
RELIABLE CHANGE DETECTION OF ARM CURL SCORES FROM BASELINE

	ROC_{auc}	PRC_{auc}	$Fscore$	$Sens.$
RF	0.58	**0.73***	**0.77***	**0.92***
SVM	0.59	**0.69***	**0.77***	**0.89***
AdaBoost	0.64	**0.76***	**0.76***	**0.84***
MLP	0.58	**0.75***	**0.69***	**0.71***

*Statistically Significant Improvement (Adjusted $p < 0.0125$) in Comparison to the Corresponding Pairwise Random Algorithm. All algorithms can build statistically significant prediction models, but the RF algorithm beats the rest in terms of $Fscore$ and Sensitivity, with similar PRC_{auc}.

TABLE VII
RELIABLE CHANGE DETECTION OF THE IMBALANCED SCORES USING SMOTEBOOST

	ROC_{auc}	PRC_{auc}	$Fscore$	$Sens.$
$RBANS_{baseline}$:				
$+\,total$	0.52	0.05	0.00	0.00
$+\,delayedmemory$	0.69	0.18	0.31	0.50
$+\,visuospatial$	0.45	0.09	0.08	0.08
$TUG_{baseline}$	0.48	0.17	0.06	0.11
$ArmCurl_{consecutive}$	0.40	0.18	0.13	0.12
$RBANS_{consecutive}$:				
$+\,delayedmemory$	0.40	0.03	0.00	0.00
$+\,visuospatial$	0.68	0.20	0.35	0.50
$TUG_{consecutive}$	**0.56***	**0.22***	**0.15***	**0.50***

*Statistically Significant Improvement (Adjusted $p < 0.005$) in Comparison to the Corresponding Pairwise Random Algorithm. Only $TUG_{consecutive}$ shows predictability.

Table VII summarizes the results for the prediction models for the imbalanced datasets that are sampled based on the SMOTEBoost algorithm. McNemar's tests for an adjusted p-value of 0.005 found significant improvement of the smart home-based prediction compared to random classifiers for the reliable change detection between consecutive assessments of TUG-based mobility. However, and even having used a method to overcome class-imbalance, models still remain biased and lacking in sensitivity.

Table VIII shows the results of the RCI detection models based on the wRACOG algorithms for the imbalanced datasets, using the sensitivity maximization as the criteria for the algorithm to stop. Compared to previous SMOTEBoost based algorithms, the sensitivity of the models is highly improved, which might be very interesting for some applications. However, some models' ROC_{auc} values lie below 0.5 and their PRC_{auc} is also low, which might again be an indicator of a biased model. In this case, the bias is towards the minority class. McNemar's tests for an adjusted p-value of 0.005 only found enough statistical significance to accept predictability of delayed memory skills between consecutive assessment points.

TABLE VIII

Reliable Change Detection of the Imbalanced Scores Using wRACOG and Sensitivity Maximization as Stopping Criteria for the Algorithm

	ROC_{auc}	PRC_{auc}	$Fscore$	$Sens.$
$RBANS_{baseline}$				
+ $total$	0.72	0.07	0.09	1.00
+ $delayedmemory$	0.63	0.10	0.13	0.60
+ $visuospatial$	0.72	0.20	0.21	1.00
$TUG_{baseline}$	0.52	0.21	0.32	0.84
$ArmCurl_{consecutive}$	0.54	0.22	0.40	0.83
$RBANS_{consecutive}$				
+ $delayedmemory$	**0.69**[*]	**0.06**[*]	**0.11**[*]	**0.80**[*]
+ $visuospatial$	0.52	0.09	0.17	1.00
$TUG_{consecutive}$	0.48	0.18	0.35	0.96

[*] Statistically Significant Improvement (Adjusted $p < 0.005$) in Comparison to the Corresponding Pairwise Random Algorithm). Only $RBANS_{baseline} - delayedmemory$ subscores show predictability.

TABLE IX

Reliable Change Detection of the Imbalanced Scores using wRACOG and ROC_{auc} Maximization as Stopping Criteria for the Algorithm

	ROC_{auc}	PRC_{auc}	$Fscore$	$Sens.$
$RBANS_{baseline}$				
+ $total$	0.77	0.07	0.17	1.00
+ $delayedmemory$	0.66	0.10	0.19	1.00
+ $visuospatial$	0.64	0.14	0.20	0.23
$TUG_{baseline}$	0.51	0.17	0.39	0.60
$ArmCurl_{consecutive}$	**0.62**[*]	**0.22**[*]	**0.49**[*]	**0.63**[*]
$RBANS_{consecutive}$				
+ $delayedmemory$	0.67	0.03	0.08	1.00
+ $visuospatial$	0.53	0.09	0.19	0.80
$TUG_{consecutive}$	0.59	0.18	0.29	0.48

[*] Statistically Significant Improvement (Adjusted $p < 0.005$) in Comparison to the Corresponding Pairwise Random Algorithm. Only $ArmCurl_{consecutive}$ shows predictability.

Table IX shows the results of the RCI detection models based on the wRACOG algorithms for the imbalanced datasets, using the ROC_{auc} metric as the stopping criteria for the iterative algorithm. The sensitivity of the models using this second approach is, overall, higher than the SMOTEBoost-based models and lower than the models presented in Table VIII. Interestingly, in some cases the areas under the ROC and PR curves, as well as the $Fscore$s, are greater than the ones obtained with the previous approaches. This suggests a better suitability of the wRACOG based models maximizing ROC_{auc} for some of the RCI detection problems. After controlling for the p-value to reduce the family-type error rate, only the model for the detection of reliable changes on consecutive Arm Curl mobility scores was showing a statistically significant prediction ability.

C. Detection of Improvement/Decline in Cognition & Mobility Skills

Table X shows the results of detecting mobility and cognition score improvement/decline. After adjusting the p-value for a reduced family-wise error rate ($^{*}p < 0.01$,$^{**}p < 0.001$), only the detection of improvement and decline in mobility as measured by the Arm Curl test seemed to be possible. A significant improvement both in ROC_{auc} and PRC_{auc} values was detected using RF and AdaBoost classifiers in comparison to their pairwise random data classifiers, as well as a significant improvement in $Fscore$ and sensitivity of the RF-based model.

IV. Discussion

The problem addressed in this work is not an easy task to solve. Our goal was to predict the multi-modal symptoms commonly seen in AD from unobtrusively-collected behavior data in smart homes with older adults residents. Despite the complexity of the task, our results show that measures of cognition, mobility, and depression are predictable using activity-labeled smart home data.

A regression analysis of the smart home-based behavior data with all the tests under analysis has shown several moderate yet significant correlations. As expected, behavior data were the most correlated to mobility assessment scores, followed by cognitive skills, whereas the most difficult task seems to be mood prediction. Nonetheless, almost all models, with the exception of cognition level prediction based on Digit Cancellation scores, showed a significant improvement compared to models based on random data.

The feature selection analysis has brought to light such valuable information as the predictability of mobility scores from outing patterns, daily routine, and patterns of cooking and eating. In the specific case of TUG scores, there was a significant correlation with global overnight activities including bed-to-toilet transitions. This finding suggests that individuals who take longer to complete the TUG (indicative of slowed movement) tend to be more active at night. This is supported by the AD literature that finds both impaired mobility and sleep disturbances to be related to dementia [40], [41]. In [12], TUG showed significant correlations with mobility, outings, sleep and ADL (cook, eat, relax and personal hygiene activities) features.

While we did not observe statistically significant predictability based on outings, mobility and sleep after adjusting the p-value for reduced family-wise error rate, we did based on global daily routine patterns, which were not analyzed previously, and for cooking and eating activities, which likely reflect part of the ADLs of the previous work. Cognition was mainly correlated to sleep and overnight patterns, but also to daily routine, mobility, and outings. These results also agree with previous work [12], where correlations between total RBANS scores and smart home activity data were analyzed and statistical significance for sleep, mobility, outings, and ADLs was found. Also in agreement with these results, sleep and sleep-related disturbances have been found to be related to cognitive impairment in other research [42], [43], as well as time spent out of home to cognitive state as measured by the Clinical Dementia Rating (CDR) scale [16].

Finally, yet lacking statistical significance for the correlation scores, depression assessed with the GDS scale was found to be predictable based on mobility, outings, and sleep features. This agrees with previous work [15] where correlation of GDS scores with overall in-home mobility and outing patterns was

TABLE X
Positive and Negative Fluctuation Detection Between Consecutive Assessment Points

	RF				SVM				AdaBoost			
	ROC_{auc}	PRC_{auc}	$Fscore$	$Sens.$	ROC_{auc}	PRC_{auc}	$Fscore$	$Sens.$	ROC_{auc}	PRC_{auc}	$Fscore$	$Sens.$
Mobility												
Arm Curl	**0.65****	**0.54****	**0.33***	**0.28***	0.60	0.38	0.38	0.34	**0.59****	**0.47****	0.36	0.36
TUG	0.41	0.49	0.38	0.39	0.46	0.45	0.45	0.48	0.45	0.49	0.46	0.50
Cognition												
PRMQ	0.54	0.47	0.29	0.25	0.56	0.38	0.39	0.35	0.51	0.45	0.41	0.43
Prospective Memory	0.58	0.44	0.26	0.21	0.50	0.31	0.19	0.16	0.58	0.42	0.35	0.37
Retrospective Memory	0.55	0.44	0.22	0.18	0.60	0.40	0.41	0.35	0.55	0.44	0.22	0.18
RBANS	0.38	0.46	0.31	0.29	0.39	0.44	0.21	0.19	0.36	0.42	0.32	0.35
Attention	0.54	0.55	0.39	0.35	0.56	0.49	0.44	0.39	0.53	0.56	0.44	0.46
Delayed Memory	0.58	0.53	0.34	0.27	0.48	0.40	0.18	0.15	0.55	0.48	0.35	0.35
Immediate Memory	0.50	0.51	0.37	0.34	0.43	0.42	0.20	0.18	0.51	0.48	0.38	0.45
Language	0.48	0.50	0.32	0.30	0.47	0.44	0.26	0.23	0.51	0.49	0.31	0.33
Visuospatial	0.48	0.51	0.32	0.30	0.57	0.52	0.43	0.39	0.43	0.48	0.35	0.36
Digit Cancel - Speed	0.44	0.45	0.28	0.25	0.51	0.43	0.36	0.32	0.43	0.43	0.32	0.34

*Statistically Significant Improvement (Adjusted $^{*}p < 0.01$, $^{**}p < 0.001$) in Comparison to the Corresponding Pairwise Random Algorithm. Only predictions of fluctuations in Arm Curl scores based on RF and AdaBoost algorithms show statistically significant predictability.

discovered. Trends showing that sleep and overnight behavior as well as daily routine features contribute most to the prediction of several health assessments are also consistent with behavior literature [42]–[45]. Thus, our results validate those reported in the literature, in addition to analyze in greater detail each aspect of mobility and cognition skills thanks to the use of more tests and their subscores. Part of the data used for these correlation analyses overlaps with the data used previously ($n = 18$) [12], so similar conclusions would be expected. Nonetheless, we have reaffirmed and given more strength to most of those conclusions by including data collected over a longer period and from more subjects (i.e., using a bigger sample size), as well as discovering new correlations with daily routine patterns. In fact, the novel overall daily-routine features presented in this paper showed predictability both for mobility and cognition skills of the elderly.

Regarding reliable change detection, we see that activity-labeled smart home data can actually be used to build quite accurate models when a complete and balanced dataset is available. This is the case for the Arm Curl test change from baseline, which has been seen to be predictable in a quite accurate manner and with a high sensitivity. We verified in all four models built for this reliable change prediction that the use of smart home activity data significantly contributes to the detection of such events. Unfortunately, a balanced dataset was not available for all cases. Despite that problem, by applying the SMOTEBoost technique to overcome class imbalance, we were able to demonstrate that consecutive reliable change on mobility measured by TUG test is predictable using smart home activity labeled data. A McNemar's test with an adjusted p-value has supported this hypothesis, yet we are aware that the model lacks sensitivity to be considered a final model. The use of the wRACOG algorithm has resulted in some models with better prediction characteristics: improved sensitivity and ROC_{auc}, PRC_{auc} and $Fscores$ were found in some cases. Changes in consecutive Arm Curl and delayed memory scores also showed enough statistical significance compared to random classifiers in a McNemar's test to be considered reliably predictable from smart home data.

Now that we know that behavioral data can be used to at least automatically assess changes in mobility and memory skills, we can keep collecting more longitudinal data to create better models in the future. This might also result in the discovery of other significant associations. Note that these results were also achieved by using all the behavioral features, whereas a feature-selection process can also help in improving them. Furthermore, we used a kNN algorithm as the wrapper model for the wRACOG approach, but other algorithms can also be considered and might improve the results. Maximization of PRC_{auc}s of $Fscore$s could also be tested as stopping criteria for the iterative process, possibly leading to different conclusions.

Analysis of the ability to detect changes in cognitive and mobility skills has demonstrated the possibility of predicting a decline or an improvement in a person's mobility as measured by the Arm Curl test. This not only confirms the results of the previous RCI analysis, where we saw that reliable changes in the Arm Curl tests were detectable by smart home activity-labeled data but also adds value to the results suggesting that the direction of the change is also predictable. Literature also supports the idea of the relationship between Arm Curl test scores and ADLs [23]. This finding may prove useful not only to monitor the progress of a disorder like dementia but also to closely examine individuals who are undergoing rehabilitation.

None of the other tests showed enough evidence of predictability after adjusting the significance level. There are several contributing factors to the difficulty of this task. On one hand, in this case, we were considering all fluctuations as labels (either positive or negative) without considering their magnitude or without taking into account their reliability (i.e., not only reliable changes were considered but all changes). This might have included "noise" in the dataset by considering changes that might have appeared due to reasons other than an actual change in the skills (such as low reliability on tests), making the classification task more difficult. On the other hand, the time-series statistics that we were extracting from the smart home behavior data do not necessarily reflect a positive or negative change in behavior, but an absolute change.

V. Conclusion

In summary, this work has demonstrated the possibility of predicting mobility, cognitive, and mood-related symptoms from unobtrusively collected in-home behavior data. We believe that the results shown herein are of high relevance, as they suggest the possibility of implementing a system that could bring huge benefits to our aging society. The models shown in this paper are early models aimed at demonstrating the feasibility of such a system and providing insight into the behavioral features that might be used for this purpose.

Completion and improvement of the results shown in this paper must be done by collecting more data and by applying algorithmic solutions that might better adapt to the imbalanced detection problems posed herein before their implementation in real-world settings. Collecting more data will also be useful to have a complete dataset with confirmed cases of transition from healthy state to cognitively impaired, which is necessary to build accurate prediction models. Thus, future work will focus on continued collection of data for further analysis, designing and testing more suitable algorithms for imbalanced datasets, and performing a more in-depth feature selection analysis in order to improve the sensitivity of the models shown herein, without the overall accuracy of the models being affected.

References

[1] M. Prince, E. Albanese, M. Guerchet, and M. Prina, "World Alzheimer Report 2014. Dementia and Risk Reduction. An analysis of protective and modifiable factors," Alzheimer's Disease Int., London, U.K., Tech. Rep. 2014. [Online]. Available: www.alz.co.uk

[2] J. Dauwels and S. Kannan, "Diagnosis of Alzheimer's disease using electric signals of the brain. A grand challenge," *Asia-Pacific Biotech News*, vol. 16, no. 10, pp. 22–38, Oct. 2012.

[3] K. G. Yiannopoulou and S. G. Papageorgiou, "Current and future treatments for Alzheimer's disease," *Therapeutic Adv. Neurol. Disorders*, vol. 6, no. 1, pp. 19–33, Jan. 2013.

[4] K. López-de Ipiña *et al.*, "On the selection of non-invasive methods based on speech analysis oriented to automatic alzheimer disease diagnosis," *Sensors*, vol. 13, no. 5, pp. 6730–6745, May 2013.

[5] A. Alberdi, A. Aztiria, and A. Basarab, "On the early diagnosis of Alzheimer's Disease from multimodal signals: A survey," *Artif. Intell. Med.*, vol. 71, pp. 1–29, Jul. 2016.

[6] R. C. Petersen, "Early diagnosis of Alzheimer's disease: Is MCI too late?" *Current Alzheimer Res.*, vol. 6, no. 4, pp. 324–30, Aug. 2009.

[7] M. Chan, D. Esteve, C. Escriba, and E. Campo, "A review of smart homes—Present state and future challenges," *Comput. Methods Programs Biomed.*, vol. 91, no. 1, pp. 55–81, 2008.

[8] P. Sajda, "Machine learning for detection and diagnosis of disease," *Annu. Rev. Biomed. Eng.*, vol. 8, pp. 537–565, 2006.

[9] E. Alpaydin, *Introduction to Machine Learning*. London, U.K.: MIT, 2010.

[10] N. J. Nilsson, *Introduction to Machine Learning*. Standford, CA, USA: Standford Univ., 1998.

[11] A. Smola and S. Vishwanathan, *Introduction to Machine Learning*. Cambridge, U.K.: Cambridge: Cambridge Univ. press, 2008.

[12] P. N. Dawadi, D. J. Cook, and M. Schmitter-Edgecombe, "Automated cognitive health assessment from smart home-based behavior data," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 4, pp. 1188–1194, Jul. 2016.

[13] T. L. Hayes, F. Abendroth, A. Adami, M. Pavel, T. A. Zitzelberger, and J. A. Kaye, "Unobtrusive assessment of activity patterns associated with mild cognitive impairment," *Alzheimer's Dementia*, vol. 4, no. 6, pp. 395–405, Nov. 2008.

[14] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen, "Mild cognitive impairment: clinical characterization and outcome," *Archives Neurol.*, vol. 56, no. 3, pp. 303–8, Mar. 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10190820

[15] C. Galambos, M. Skubic, S. Wang, and M. Rantz, "Management of dementia and depression utilizing in-home passive sensor data," *Gerontechnology*, vol. 11, no. 3, pp. 457–468, 2013.

[16] J. Petersen, D. Austin, N. Mattek, and J. Kaye, "Time out-of-home and cognitive, physical, and emotional wellbeing of older adults: A longitudinal mixed effects model," *PLoS ONE*, vol. 10, no. 10, pp. 1–16, 2015.

[17] J. Austin, H. H. Dodge, T. Riley, P. G. Jacobs, S. Thielke, and J. Kaye, "A smart-home system to unobtrusively and continuously assess loneliness in older adults," *IEEE J. Translational Eng. Health Med.*, vol. 4, pp. 1–11, 2016.

[18] L. Christensen and J. L. Mendoza, "A method of assessing change in a single subject: An alteration of the RC index," *Behavior Therapy*, vol. 17, no. 3, pp. 305–308, Jun. 1986.

[19] P. Dawadi, D. J. Cook, and M. Schmitter-Edgecombe, "Automated clinical assessment from smart home-based behavior data," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 4, pp. 1188–1194, Jul. 2016.

[20] C. Randolph, M. C. Tierney, E. Mohr, and T. N. Chase, "The repeatable battery for the assessment of neuropsychological status (RBANS): Preliminary clinical validity," *J. Clin. Exp. Neuropsychol. (Neuropsychol., Develop. Cognition, Sect. A)*, vol. 20, no. 3, pp. 310–319, Jun. 1998.

[21] J. Crawford, G. Smith, E. Maylor, S. Della Sala, and R. Logie, "The prospective and retrospective memory questionnaire (PRMQ): Normative data and latent structure in a large non-clinical sample," *Memory*, vol. 11, no. 3, pp. 261–275, Jan. 2003.

[22] D. Podsiadlo and S. Richardson, "The timed "Up & Go": A test of basic functional mobility for frail elderly persons," *J. Am. Geriatrics Soc.*, vol. 39, no. 2, pp. 142–148, Feb. 1991.

[23] R. E. Rikli and C. J. Jones, "Development and validation of a functional fitness test for community-residing older adults," *J. Aging Physical Activity*, vol. 7, no. 2, pp. 129–161, Apr. 1999.

[24] T. Hatta, K. Yoshizaki, Y. Ito, M. Mase, and H. Kabasawa, "Reliability and validity of the digit cancellation test, a brief screen of attention," *Psychologia*, vol. 55, no. 4, pp. 246–256, 2012.

[25] J. I. Sheikh and J. A. Yesavage, "Geriatric depression scale (GDS) recent evidence and development of a shorter version," *Clin. Gerontol.*, vol. 5, no. 1/2, pp. 165–173, Nov. 1986.

[26] E. Smith, L. Walsh, J. Doyle, B. Greene, and C. Blake, "The reliability of the quantitative timed up and go test (QTUG) measured over five consecutive days under single and dual-task conditions in community dwelling older adults," *Gait Posture*, vol. 43, pp. 239–244, Jan. 2016.

[27] J. M. Miotto, W. J. Chodzko-Zajko, J. L. Reich, and M. M. Supler, "Reliability and validity of the fullerton functional fitness Test: An independent replication study," *J. Aging Phys. Activity*, vol. 7, no. 4, pp. 339–353, Oct. 1999.

[28] M. C. Salinsky, D. Storzbach, C. B. Dodrill, and L. M. Binder, "Test-retest bias, reliability, and regression equations for neuropsychological measures repeated over a 12-16-week period," *J. Int. Neuropsychol. Soc.*, vol. 7, no. 5, pp. 597–605, Jul. 2001.

[29] C. McKay, J. Casey, J. Wertheimer, and N. Fichtenberg, "Reliability and validity of the RBANS in a traumatic brain injured sample," *Archives Clin. Neuropsychol.*, vol. 22, no. 1, pp. 91–98, Jan. 2007.

[30] J. Crawford, G. Smith, E. Maylor, S. Della Sala, and R. Logie, "The prospective and retrospective memory questionnaire (PRMQ): Normative data and latent structure in a large non-clinical sample," *Memory*, vol. 11, no. 3, pp. 261–275, Jan. 2003.

[31] O. Pedraza, V. M. Dotson, F. B. Willis, N. R. Graff-Radford, and J. A. Lucas, "Internal consistency and test-retest stability of the geriatric depression scale-short form in African American older adults," *J. Psychopathol. Behavioral Assessment*, vol. 31, no. 4, pp. 412–416, Dec. 2009.

[32] N. C. Krishnan and D. J. Cook, "Activity recognition on streaming sensor data," *Pervasive Mobile Comput.*, vol. 10, no. PART B, pp. 138–154, 2014.

[33] J. W. Ratcliff and D. E. Metzener, *Pattern-Matching-the Gestalt Approach*. San Mateo, CA, USA: Miller Freeman, 1988, vol. 13, no. 7.

[34] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques."* 4th ed. San Mateo, CA, USA: Kaufmann, 2016.

[35] F. Segovia, C. Bastin, E. Salmon, J. M. Górriz, J. Ramirez, and C. Phillips, "Automatic differentiation between Alzheimer's Disease and mild cognitive impairment combining PET data and psychological scores," in *Proc. 2013 Int. Workshop Pattern Recognit. Neuroimag.*, 2013, pp. 144–147.

[36] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," in *Proc. Eur. Conf. Principles Data Mining Knowl. Discovery*, Dec. 2003, pp. 107–119.

[37] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Int. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[38] B. Das, N. C. Krishnan, and D. J. Cook, "WRACOG: A Gibbs sampling-based oversampling technique," *Proc. IEEE Int. Conf. Data Mining*, pp. 111–120, 2013.

[39] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 462–467, May 1968.

[40] M. I. Tolea, J. C. Morris, and J. E. Galvin, "Trajectory of mobility decline by type of dementia," *Alzheimer Disease Associated Disorders*, vol. 30, no. 1, pp. 60–66, 2016.

[41] S. Ooms and Y.-E. Ju, "Treatment of sleep disorders in dementia," *Current Treatment Options Neurol.*, vol. 18, no. 9, p. 40, Sep. 2016.

[42] T. L. Hayes, T. Riley, N. Mattek, M. Pavel, and J. A. Kaye, "Sleep habits in mild cognitive impairment," *Alzheimer Disease Assoc. Disorders*, vol. 28, no. 2, pp. 145–150, 2014.

[43] R. A. P. C. da Silva, "Sleep disturbances and mild cognitive impairment: A review," *Sleep Sci.*, vol. 8, no. 1, pp. 36–41, Jan. 2015.

[44] D. E. Vance, K. Heaton, Y. Eaves, and P. L. Fazeli, "Sleep and cognition on everyday functioning in older adults," *J. Neurosci. Nursing*, vol. 43, no. 5, pp. 261–271, Oct. 2011.

[45] V. Bergua, C. Fabrigoule, P. Barberger-Gateau, J.-F. Dartigues, J. Swendsen, and J. Bouisson, "Preferences for routines in older people: associations with cognitive and psychological vulnerability," *Int. J. Geriatric Psychiatry*, vol. 21, no. 10, pp. 990–998, Oct. 2006.