Smart Homes predicting the Multi-Domain Symptoms of Alzheimer's Disease

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

Abstract—As members of an increasingly aging society, one of our major priorities is the development of tools to early detect 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 behavioral data to detect the multimodal symptoms of AD. After gathering longitudinal smart home data of 29 older adults for an average of >2 years, we automatically labeled the data with corresponding activity classes and extracted time-series statistics containing 10 behavioral features. AD symptoms were assessed every six months by means of self-reported mobility, memory/cognition and mood tests. Using these data, we created regression models to predict AD symptoms as measured by the tests and a feature selection analysis was performed. Classification models were built to detect reliable absolute change in the scores predicting AD and SmoteBOOST algorithm was used to overcome class imbalance where needed. Results show that all mobility, cognition/memory and depression symptoms are predictable by the activity-aware smart home data, as well as a reliable change in mobility and visuospatial skills related to cognition. Results also suggest that not all behavioral features contribute equally to the prediction of every symptom. Future work must focus on improving the sensitivity of the presented models by collecting more longitudinal data and by focusing on class-imbalance suitable algorithms and in in-depth feature selection. The results presented herein contribute significantly towards the development of an early AD detection system based on smart home technology.

Index Terms—Activity Recognition, Alzheimer's Disease, Automatic Assessment, Behavior, Multimodal Symptoms, Older Adults, Smart Home.

1 Introduction

The increasing life expectancy in the developed countries has resulted in more and more 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], and not having a definitive cure for it yet [2], this can result in devastating consequences in terms of health-care costs and quality of life of patients and relatives. As a matter of general interest, the search for solutions for the early detection and cure for AD is currently a high priority issue.

AD manifests symptoms in multiple domains, such as psychology, physiology, behavior and cognition [3]. These symptoms are usually measured by means of self and informant-reported tests, by interviews with psychologists or physicians or in costly medical examinations based on brain imaging, which are often performed too late, resulting in a delayed diagnosis. Nowadays, only treatments to relieve AD's cognitive and behavioral symptoms are available [4], but the key for even these to be effective is the early detection of the disease.

Smart homes are an emerging technological solution enabling the monitoring of people's behavior unobtrusively and ubiquitously [5]. Real-life data can be gathered non-stop throughout the day in a completely transparent way for the user, offering a complete view of older adults' behavior and allowing the detection of changes that might indicate the onset of a disorder at all times. If smart home-based behavior shifts were mapped to AD, the main disadvantages of the usual assessment methods could be overcome making an early diagnosis of the disorder possible.

Our goal in this paper is to assess the possibility of detecting the psychological, cognitive and behavioral symptoms of AD making use of unobtrusively collected smart home behavior data. The affirmation of this hypothesis would result in a path to follow towards the final development and implementation of an early AD detection system that could alert patients and relatives on time gaining efficacy of treatments.

The main contributions of this work can be summarized as follows. 1) The predictability of wide variety of health assessment scales measuring AD's multi-domain symptoms or its onset is analyzed for the first time. 2) A throughout analysis about the contribution of each behavioral feature to the prediction of each health assessment score is done. 3) Some new smart home-based behavior features aiming at measuring the global daily routine of the elderly are presented and their contribution to the scales under analysis is evaluated. 4) Finally, the problem of detecting a reliable change in the health assessment scores of the elderly from unobtrusively collected behavioral data is addressed using

- *A. Alberdi, A. Aztiria and M. Barrenechea are with the Electronics and Computing Department, Goiru Kalea, 2, Arrasate, 20500, Spain E-mail: aalberdiar,aaztiria,mbarrenetxea@mondragon.edu*
- *A. Weakley and Maureen Schmitter-Edgecombe are with Department of Psychology, Washington State University, Pullman, Washington, 99164, USA. E-mail: alymae,schmitter-e@wsu.edu*
- *D. J. Cook is with School of Electrical Engineering and Computer Science, Washington State University,Pullman, Washington, 99164, USA E-mail: djcook@wsu.edu*
- *A. Basarab is with Université de Toulouse, Institut de Recherche en Informatique de Toulouse, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5505, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse, France E-mail: adrian.basarab@irit.fr*

Manuscript received X X, X; revised X X, X.

specific algorithms (SmoteBOOST) to overcome the so common class imbalance of health-related research.

This paper is structured as follows: In section 2 the related literature is reviewed. In section 3, the steps followed to collect and process the data, as well as to create and validate the prediction models is explained. Section 4 summarizes the results of the study, and finally in section 5 the main conclusions and future lines are stated.

2 Related Work

Previous research has demonstrated that longitudinal monitoring of smart home-based behavioral data can be useful not only to assess older adults' health state but also to detect the onset and follow the progress of some age-related diseases and disorders. Dawadi *et al.* affirmed that the overall cognitive and mobility skills of older adults can be predicted by unobtrusively collected in-home behavioral data [6]. For that purpose, they introduced an algorithm called Clincal Assessment using Activity Behavior (CAAB) and tested its validity for global Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) and Timed Up and Go (TUG) scores' prediction using time-series statistics of several activities of daily living as predictors. Hayes *et al.* have found MCI [7] 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, whereas Galambos *et al.* [8] discovered associations between overall in-home activity and outings patterns with both dementia and depression, which is also known to be a common AD symptom. MMSE, Short Form Health Survey-12 and GDS scales were used to determine subjects' state. Petersen *et al.* [9] also affirmed emotional states in terms of 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. Loneliness of older adults has also been predicted by analyzing their behavioral data by Austin et al [10] .

3 Methods

3.1 Data collection

First, we accessed unobtrusively collected in-home behavioral data of 40 older adults living in 38 Smart-Homes, as well as their biannual neuropsychological assessment data, which were collected by the Center for Studies in Adaptive Systems (CASAS) and the Neuropsychology and Aging Laboratory at Washington State University (WA, USA).

Neuropsychological assessments of the subjects were performed twice a year. Even if a wide variety of overall health and psychology checking tests were performed in each one of the assessment sessions, for the current study we'll focus on cognition/memory, mobility and mood (depression) scores (see table 1), which have been found to be affected by AD [3]. Cognitive abilities of the older adults were measured by means of the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) [11], the Prospective and Retrospective Memory Questionnaire (PRMQ) [12] and a Digit Cancellation test, while mobility was assessed by Timed Up and Go (TUG) [13] and Arm Curl [14] tests. The Geriatric Depression Scale - Short Form (GDS-15) [15] was used to assess the depression level of the elderly under study.

TABLE 1: 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	[16]
	TUG	0.96	3.18	[17]
Cognition / Memory	RBANS - total	0.88	14.04	[18]
	RBANS - attention	0.88	17.47	
	RBANS - delayed memory	0.88	19.62	
	RBANS - immediate memory	0.88	17.12	
	RBANS - visuospatial	0.88	12.96	
	RBANS - language	0.88	15.60	
	PRMQ - total	0.89	9.15	[19]
	PRMQ - prospective memory	0.85	4.91	
	PRMQ - retrospective memory	0.89	4.98	
	Digit Cancel	0.85	37.20	[20]
Mood	GDS	0.68	2.20	[21]

Smart home sensor data collection started in 2011 and lasted until 2016, a period in which the data were collected continuously for lengths ranging from $<$ 1 month to 60 months (M= 19.95 months, SD=17.98 months) depending on each apartment. As data coming from homes with multiple habitants might pose problems to correctly estimate each individual's activity level, these were removed for the following analyses. Subjects with missing health assessment data or with behavioral data collected for less than 6 months were also removed. Hence, the final dataset contained the behavioral and health assessment data of multiple modalities (cognition/memory, mobility and mood) of 29 older adults who were living independently and alone in their own smart home residences (M=26 months, SD=17.5 months, range=6-60 months).

3.2 Preprocessing

3.2.1 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 raw-sensor data stream was an entry specifying the events' timestamp, ID of the sensor detecting the event and type of event (activation/deactivation). In order to make the raw sensor-data streams interpretable, it was first necessary to assign a specific activity to each sensor entry. For that purpose, the AR activity recognition algorithm specified in [22] was used. The reliability of this algorithm has been demonstrated in a previous work, where accuracy greater than 98% was achieved on 30 testbed smart homes using three-fold cross validation.

Once the activity-level information was available, we computed 17 daily behavior features for each subject, explaining their daily sleep and mobility patterns, time spent in some specific ADLs and overall characteristics of their routines. A detailed list of the computed features can be seen in table 2.

The daily distance that the subjects were traveling inside their homes was estimated by creating sensor mapping-files based on the floor plan and sensor layout for each residence (see example in Figure 1), where the x-y coordinates of the motion sensor's positions were specified. Three of the apartments lacked specific information about the positioning of

TABLE 2: 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, carrying out personal hygiene activities, being out of home and nighttime toileting activities
Sleep-related (2 features)	The daily sleep duration and frequency
Mobility-related (2 features)	The total number of activated sensors and the total distance covered walking inside the apartment 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 previous day

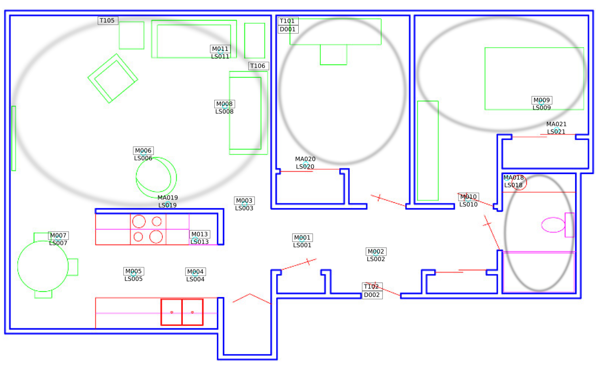

Fig. 1: Floor plan and sensor layout of one of the residences of the study

the sensors and/or the distribution within the houses. In those cases, it was first necessary to estimate the positions of the sensors, which was done by considering the 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 doesn't take into account the existence of walls or other obstacles between the sensors that must be avoided or surrounded.

To compute daily-routine features, we first extracted the daily activity sequence from the AR-labeled sensor data stream. We then encoded the daily activity sequence by replacing each activity with a number from 1 to 12 (i.e. Sleep=1, Cook=2, Relax=3, ..., Others = 12). Shannon entropy was used as the measure of complexity of the daily routine. So as to get this entropy value, we computed the daily probability distribution (histogram) of the activity sequence (P), and we then applied the Shannon entropy as in (1),

$$Complexity_{routine} = \sum_{activity=1}^{12} P_{activity} - log_2 P_{activity} \quad (1)$$

where $P_{activity}$ was the probability of a certain activity to happen during the 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 [23] available in Python as "SequenceMatcher" function, which expresses the similarity of any two sequences as a ratio between 0 and 1. This allowed us to measure 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 from the first to the last detected activity of the day. We believe that the remaining features are self-explanatory.

3.2.2 Between-assessments behavior statistics' computation

At the end of the previous step, we had available a set of daily activity features for each subject. We then applied the CAAB algorithm, which was introduced in [22], to the daily activity data in order to extract the behavioral statistics of each between-assessment period. RStudio for R was the selected environment for this purpose.

In short, 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 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 4% and 4) Compute the average of each time-series statistic for the 6-month period and use them 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 of 31 older adults who were living alone in their sensorized apartments for a period of $24.0 \pm 13.68(SD)$ months.

3.2.3 Health assessment scores' set-up

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 and memory assessment based on Digit-Cancel test and RBANS and PRMQ score and subscores, as well as depression symptoms represented as GDS test-scores. All these values were self-reported by the participants at the end of each corresponding 6-month period.

Self-reported questionnaires can be highly subject-dependent for several reasons. In order to take into account the inter-subject variability that each subject's age, gender, education or habits might provoke in the scores, we also considered to standardize them for each one of the subjects by computing a Reliable Change Index (RCI) [24] that informs whether a subject has suffered a significant change in an assessment score based on his/her own previous performance. The RCI discards changes that might have appeared due to reasons other than an actual change in scores (such as measurement unreliability) by applying a threshold to the scores' differences. We looked for both reliable absolute changes compared to baseline values ($RCI_{baseline}$) and compared to the previous assessment point ($RCI_{consecutive}$) of each subject for all tests' outputs.

In order to calculate the RCIs for the scores used herein, we gathered test-retest reliability (r_{score}) and standard deviations (SD_{score}) that the tests have shown in their development cohorts and/or in previous works, as shown in table 1. 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.

Nonetheless, a Reliable change in a health-assessment score is a rare event. As such, we had very few positive instances (data instances where a reliable change was observed) for some of the assessment scores, resulting in highly imbalanced data. For the following, we removed from the study those tests which were extremely imbalanced (<10% of positive instances). We distinguished the remaining tests in imbalanced (10%-30% of positive instances) and not-imbalanced data (30%-50% of positive instances).

3.3 Cognition and mobility change prediction

The preprocessed dataset resulting from the previous steps was analyzed using Weka.

First, we performed a correlation analysis between the mobility, cognition/memory and mood assessment scores and smart home behavior data. For this purpose, we implemented four different regression models using all behavioral features computed in the previous step for each one of the scores using Support Vector Regression (SVr) with a linear kernel, Linear Regression (LinearR), SVr with a Radial Basis Function (RBF) kernel and k nearest neighbours (kNN) algorithms. We compared the correlation coefficients (r) and Mean Absolute Errors (MAE) of the models using 10-fold Cross-Validation 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 significant improvement of smart home-based algorithms in comparison to the random data algorithms.

In order to analyze the types of behavioral features that are most correlated with each one of the tests, we built and evaluated activity-specific models for the main test scores with a 10-fold CV. The behavioral features that were included in each one of the models are shown in table 3.

TABLE 3: Task-specific grouping of the 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 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

Again, we searched for statistically significant improvement in comparison to pairwise random algorithms using a corrected paired t-test.

Regarding RCI detection, we used different approaches for the imbalanced and not-imbalanced datasets. First, not imbalanced datasets containing all behavioral features were reduced by means of a Principal Component Analysis. Principal Components explaining the 95% of the variability in the behavior data were kept to create the reduced datasets. SVM, AdaBoost, Multilayer Perceptron (MLP) and Random Forest (RF) algorithms were trained and validated following a 10-fold CV approach. Area under the ROC curve (AUC ROC), area under the Precision-Recall curve (AUC PRC), F-score and sensitivity were selected as the metrics for models' evaluation. The combination of these metrics offers an excellent overview of both models' overall performance and capability to detect the event of interest (the reliable change event), and are specially suitable when the data to work with is skewed. A corrected paired t-test was used to detect significant improvement of smart home-based algorithms in comparison to the pairwise random data algorithms.

For the imbalanced datasets, a different approach was required. Common machine-learning algorithms tend to create biased models 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 event in which we are more interested is the rare event or the minority class, highlighting the need to use alternative methods to improve the detection of these minority events. SMOTEBoost [25] is a method combining Boosting techniques with SMOTE [26] oversampling techniques. Whereas boosting aims at creating a "strong" classifier using a set of "weak" classifiers, SMOTE is a technique to oversample the minority class by creating synthetic data instances and thus, reduce class imbalance. SMOTEBoost combines these processes iteratively in order to improve the sensitivity of the models without the overall accuracy being affected.

We built prediction models for imbalanced datasets using SMOTEBoost and kNN with k=5 as the "weak" classifier. A 3-fold CV was performed for validation purposes. 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, AUC ROC, AUC PRC, F-score and sensitivity of the models were computed for models' performance screening. McNemar's test was applied to check whether a significant improvement was observed using smart home data in the

TABLE 4: Regression results for the absolute test scores using all behavioral features for 10-fold CV. (Statistically significant improvement (*p<0.05,**p<0.01,***p<0.001) in comparison to the corresponding pairwise random algorithm.))

	SVR		LinearR		SVR RBF		kNN	
	r	MAE	r	MAE	r	MAE	r	MAE
Mobility								
Arm Curl	0.17	5.62	0.06	4.99	0.29**	4.18*	0.14	5.91
TUG	0.51***	3.89	0.36***	3.92	0.57***	3.03***	0.42***	3.72
Cognition & Memory								
PRMQ	0.26**	9.28	0.20*	8.42	0.31***	7.01	0.31**	8.40
Prospective Memory	0.30**	4.72	0.12	4.55	0.26**	3.72	0.28**	4.50
Retrospective Memory	0.15	5.15	0.14	5.17	0.39***	3.57**	0.27*	4.53
RBANS	0.27*	15.97	0.34***	15.50	0.40***	13.07*	0.34***	15.74
Attention	0.40***	16.14	0.33***	17.12	0.31***	15.52	0.22*	16.80
Delayed Memory	0.13	15.01	0.04	17.65	0.31**	11.19	0.31**	16.39
Immediate Memory	0.00	16.23	0.03	15.43	0.08	12.04	0.29***	14.52
Language	0.47***	10.77	0.35***	13.25	0.47***	10.10**	0.26**	12.41
Visuospatial	0.01	18.25	0.04	17.51	0.21**	12.28	0.18	15.12
Digit Cancel - Speed	0.22*	35.05	0.17	46.10	0.18	26.88	0.23*	31.26
Mood / Depression								
GDS	0.02	2.79	0.02	2.84	0.21*	1.67**	0.12	1.97

prediction of reliable change in the scores in comparison to random data algorithms.

4 RESULTS

4.1 Absolute test scores' prediction

Table 4 shows the results of the regressions for all the absolute test scores using all smart home behavioral features. For mobility tests, whereas Arm Curl has a weak correlation with behavioral data, TUG has been found to be strongly correlated with behavioral data. Regarding cognition and memory overall scores and subscores, all of them show moderate correlations with behavioral data except the visuospatial and immediate memory subscores of the RBANS test, which were found to correlate weakly. Finally, depression has shown a weak correlation with the global set of smart home behavioral data.

Regressions based on specific activities, which can be seen in table 5 have shown some interesting results. Arm Curl mobility test has shown weak but, statistically significant correlations only with outings, daily routine, cooking and eating and mobility features. In contrast, TUG test has shown significant moderate correlations with daily routine, sleep, overnight toileting and the combination of the last two, as well as a significant weak correlation with cooking and eating features.

Regarding cognition and memory tests, global PRMQ score was moderately associated to daily routine and to the overnight patterns, as well as weakly correlated to sleep, overnight toileting and mobility. RBANS was moderately correlated with overnight patterns, whereas it was also showing weak yet statistically significant correlations with mobility, mobility and outings, sleep and overnight toileting behaviors. Digit Cancel processing speed was found to be moderately correlated to sleep and overnight patterns, and weakly yet significantly to overnight toileting, daily routine and cooking and eating features.

Finally, for the geriatric depression assessment, we've found weak yet significant correlations with mobility alone, mobility and outings and sleep features.

TABLE 5: Regression results for the absolute test scores by behavioral feature type for 10-fold CV. (Statistically significant improvement (*p<0.05,**p<0.01,***p<0.001) in comparison to the corresponding pairwise random algorithm.))

(a) Mobility - Arm Curl

	SVr		LinearR		SVr RBF		kNN	
	r	MAE	r	MAE	r	MAE	r	MAE
Daily routine	0.28*	4.35	0.17	4.83	0.16	4.43	0.05	6.52
Mobility	0.09	4.93	0.12*	4.71	0.02	4.61	0.18	5.80
Mobility & outings	0.14	4.58	0.11	5.01	0.11	4.42	0.15	5.79
Outings	0.28**	4.30	0.20	4.46	0.23*	4.40	0.00	6.45
Sleep	0.17	4.99	0.11	4.74	0.00	4.57	0.17	5.77
Overnight patterns	0.03	4.81	0.16	4.63	0.10	4.65	0.04	6.51
Overnight toileting	0.10	4.67	0.14	4.66	0.06	4.57	0.12	7.21
Cook & eat	0.06	4.85	0.20*	4.55	0.27**	4.32	0.08	6.19

(b) Mobility - TUG

	SVr		LinearR		SVr RBF		kNN	
	r	MAE	r	MAE	r	MAE	r	MAE
Daily routine	0.35***	3.77	0.33**	3.97	0.37***	3.52**	0.32*	3.77
Mobility	0.12	3.74	0.04	4.21	0.14	3.68*	0.08	4.71
Mobility & outings	0.18	3.74	0.11	4.19	0.16	3.70*	0.10	4.61
Outings	0.00	3.96	0.07	4.08	0.08	3.76	0.08	4.73
Sleep	0.30*	3.66	0.30*	3.75	0.25*	3.71	0.13	4.53
Overnight patterns	0.33**	3.59	0.30*	3.78	0.32***	3.53*	0.19	4.21
Overnight toileting	0.26**	3.54**	0.30**	3.81	0.29**	3.50**	0.14	4.62
Cook & eat	0.16*	3.78	0.24**	3.91	0.10	3.73	0.19	5.92

(c) Cognition & Memory - PRMQ

	SVr		LinearR		SVr RBF		kNN	
	r	MAE	r	MAE	r	MAE	r	MAE
Daily routine	0.18	7.96	0.20*	7.93	0.21*	7.42	0.32**	8.15
Mobility	0.12	7.84	0.21*	7.45	0.16	7.35	0.01	10.63
Mobility & outings	0.09	8.03	0.14	7.91	0.14	7.44	0.17	10.28
Outings	0.05	7.74	0.10	7.88	0.04	7.59	0.19	8.73
Sleep	0.28**	7.13	0.19	7.56	0.24*	7.57	0.11	9.86
Overnight patterns	0.30**	7.37	0.27*	7.42	0.29**	7.47	0.25*	8.51
Overnight toileting	0.29**	7.18	0.23*	7.43	0.25*	7.54	0.28***	8.13
Cook & eat	0.11	7.65	0.07	7.92	0.09	7.54	0.01	12.29

(d) Cognition & Memory - RBANS

	SVr		LinearR		SVr RBF		kNN	
	r	MAE	r	MAE	r	MAE	r	MAE
Daily routine	0.13	15.64	0.13	16.43	0.25**	14.27	0.18	17.86
Mobility	0.18*	14.51	0.23*	14.59	0.26**	14.21	0.18*	19.00
Mobility & outings	0.14	15.24	0.22*	14.73	0.25*	14.11	0.15	17.69
Outings	0.01	14.62	0.09	14.6	0.01	14.45	0.16	16.37
Sleep	0.20	14.36	0.20	14.72	0.24*	14.12	0.19	17.12
Overnight patterns	0.30**	14.18	0.31**	13.46	0.31**	13.61*	0.17	17.17
Overnight toileting	0.26**	13.66	0.22*	13.91	0.25**	13.72**	0.03	18.61
Cook & eat	0.02	14.78	0.10	14.72	0.04	14.23	0.02	18.83

(e) Cognition & Memory - Digit Cancel

	SVr		LinearR		SVr RBF		kNN	
	r	MAE	r	MAE	r	MAE	r	MAE
Daily routine	0.23*	26.86	0.20*	32.44	0.16	27.08	0.05	37.28
Mobility	0.07	27.77	0.07	29.04	0.08	27.33	0.03	36.70
Mobility & outings	0.00	29.35	0.13	31.26	0.05	27.62	0.13	33.43
Outings	0.12	28.94	0.11	30.22	0.06	27.55	0.02	36.84
Sleep	0.22	26.98	0.30**	26.44	0.20	26.92	0.04	37.91
Overnight patterns	0.22*	28.52	0.31**	28.03	0.20	27.10	0.09	33.83
Overnight toileting	0.22*	27.49	0.29**	27.68	0.14	27.12	0.17	31.81
Cook & eat	0.21*	30.06	0.17*	31.59	0.18	27.56	0.02	46.91

(f) Mood - GDS

	SVr		LinearR		SVr RBF		kNN	
	r	MAE	r	MAE	r	MAE	r	MAE
Daily routine	0.16	1.85	0.15	2.38	0.16	1.67*	0.07	1.92
Mobility	0.12	1.72*	0.25*	1.9	0.19	1.66**	0.13	1.97
Mobility & outings	0.21*	1.68**	0.24*	1.95	0.25*	1.62**	0.07	2.07
Outings	0.14	1.73	0.17	1.98	0.19	1.67*	0.03	2.55
Sleep	0.26*	1.67**	0.21*	1.97	0.25*	1.66**	0.08	2.07
Overnight patterns	0.19	1.77	0.13	2.07	0.22	1.67*	0.06	2.05
Overnight toileting	0.02	1.76	0.08	1.91	0.09	1.74	0.14	2.01
Cook & eat	0.08	1.83	0.04	2.07	0.09	1.70*	0.01	2.24

TABLE 6: Reliable change detection of Arm Curl scores from baseline (*: Statistically significant improvement (adjusted p<0.0125) in comparison to the corresponding pairwise random algorithm.)

	AUC ROC	AUC PRC	F-score	Sensitivity
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***

TABLE 7: Reliable change detection of the imbalanced scores using SMOTEBoost (*: Statistically significant improvement (adjusted p<0.005) in comparison to the corresponding pairwise random algorithm.)

	AUC ROC	AUC PRC	F-score	Sensitivity
$RBANS_{baseline} - total$	0.66	0.17	0.09	0.13
$RBANS_{baseline} - language$	0.55	0.19	0.18	0.21
$RBANS_{baseline} - visuospatial$	0.60	0.21	0.25	0.21
$TUG_{baseline}$	0.48	0.17	0.063	0.11
$ArmCurl_{consecutive}$	0.40	0.18	0.13	0.12
$RBANS_{consecutive} - language$	0.54	0.11	0.00	0.00
$RBANS_{consecutive} - visuospatial$	**0.57***	**0.24***	**0.29***	**0.30***
$TUG_{consecutive}$	**0.56***	**0.22***	**0.15***	**0.50***

4.2 RCI detection

Global PRMQ and subscores, consecutive global RBANS scores, RBANS subscores related to attention, delayed memory and immediate memory, Digit cancel score and GDS test were excluded from the study as they were capturing less than a 10% of reliable change instances. Among the remaining labels, only the reliable change in Arm Curl score from baseline had enough positive instances to be considered a balanced dataset. The remaining scores (RBANS, RBANS language, RBANS visuospatial and TUG change from baseline, and RBANS language, RBANS visouspatial and TUG change between consecutive assessments) were considered imbalanced and processed as such.

Table 6 shows the results for Arm Curl reliable change detection from baseline. All four classifiers show a statistically significant improvement in terms of Area under the PR curve, F-score and sensitivity for the adjusted p-value, whereas area under the ROC curve shows reasonable results surpassing the 0.6 barrier.

Table 7 summarized the results for the prediction models for the imbalanced datasets based on SMOTEBoost algorithm. McNemar's tests have found significant improvement of the smart home based prediction models compared to random classifiers for an adjusted p-value of 0.005 for the reliable change detection between consecutive assessments in visuospatial skills measured by RBANS test and mobility measured by the TUG test. However, and even having used methods to overcome class-imbalance, models remain yet biased and lacking sensitivity.

5 Conclusion

The problem addressed in this work is not an easy task to solve. Our goal was to predict the multi-modal symptoms of AD from unobtrusively collected behavioral data inside older adults' apartments. Despite the complexity of the task, our results show that AD symptoms related to 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 have shown several significant correlations. As expected, behavioral data were the most correlated to mobility assessment scores, followed by cognitive and memory skills, whereas the most difficult task seems to be mood prediction. Nonetheless, all models have shown 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, cooking and eating patterns and mobility features. In the specific case of TUG score there was also a significant correlation with sleep & overnight patterns. In [6], TUG also showed significant correlations with mobility, outings, sleep and ADL features. Memory and cognition were mainly correlated to sleep and overnight patterns, but also to daily routine, mobility, outings and cooking and eating features. These results also agree with previous work [6], where correlations between total RBANS score and smart home activity data were analyzed. Finally, depression assessed with the GDS scale was found to correlate with mobility, outings and sleep features. This agrees with previous work [8] where correlation of GDS score with overall in-home mobility and outing patterns was discovered. Thus, our results validate those reported in the literature, in addition to analyzing more in detail each aspect of mobility and memory/cognition skills thanks to the use of more tests and their subscores, as well as discovering new correlations with daily routine patterns.

Regarding reliable change detection, we've seen 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 of Arm Curl test change from baseline, which has been seen to be predictable in a quite accurate manner and with a high sensitivity. We have 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, we didn't have a balanced dataset available for all cases. Despite that problem, by applying SMOTEBoost technique to overcome class imbalance, we were able to demonstrate that consecutive reliable change on mobility measured by TUG test and consecutive reliable change on visuospatial abilities measured by RBANS test are 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 models built in this work lack sensitivity to be considered final models. Now that we know that behavioral data can be used to at least automatically assess changes in mobility and visuospatial 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.

Summing up, this work has demonstrated the possibility of predicting AD mobility, cognitive and mood-related symptoms from unobtrusively collected in-home behavioral data. We believe that the results shown herein are of high relevance in our increasingly older society, which is suffering more and more AD incidence, as they suggest the possibility of implementing a system that could bring huge benefits to it. 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. Thus, future work will focus on keeping collecting data for further analysis, on testing or designing more suitable algorithms for imbalanced datasets an 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.

Acknowledgments

Research reported in this paper was supported by the National Institutes of Health under award number R01EB015853.

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 International (ADI), London, 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. 10n11, pp. 22–38, oct 2012. [Online]. Available: http://www.asiabiotech.com/publication/apbn/16/english/preserved-docs/1610n11/1610n11.pdf

[3] A. Alberdi, A. Aztiria, and A. Basarab, "On the early diagnosis of Alzheimer's Disease from multimodal signals: A survey," *Artificial Intelligence in Medicine*, vol. 71, pp. 1–29, jul 2016. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0933365716300732

[4] K. G. Yiannopoulou and S. G. Papageorgiou, "Current and future treatments for Alzheimer's disease," *Therapeutic Advances in Neurological Disorders*, vol. 6, no. 1, pp. 19–33, jan 2013. [Online]. Available: http://tan.sagepub.com/cgi/doi/10.1177/1756285612461679

[5] M. Chan, D. Esteve, C. Escriba, and E. Campo, "A review of smart homes - Present state and future challenges," *Computer Methods and Programs in Biomedicine*, vol. 91, no. 1, pp. 55–81, 2008. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0169260708000436

[6] P. N. Dawadi, D. J. Cook, and M. Schmitter-Edgecombe, "Automated Cognitive Health Assessment From Smart Home-Based Behavior Data," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 4, pp. 1188–1194, jul 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7206522/

[7] 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. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1552526008028653

[8] 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.

[9] 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.

[10] 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 Journal of Translational Engineering in Health and Medicine*, vol. 4, pp. 1–11, 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7488979/

[11] C. Randolph, M. C. Tierney, E. Mohr, and T. N. Chase, "The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): Preliminary Clinical Validity," *Journal of Clinical and Experimental Neuropsychology (Neuropsychology, Development and Cognition: Section A)*, vol. 20, no. 3, pp. 310–319, jun 1998. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1076/jcen.20.3.310.823

[12] 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. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/09658210244000027

[13] 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*, vol. 39, no. 2, pp. 142–8, feb 1991. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/1991946

[14] R. E. Rikli and C. J. Jones, "Development and Validation of a Functional Fitness Test for Community-Residing Older Adults," *Journal of Aging and Physical Activity*, vol. 7, no. 2, pp. 129–161, apr 1999. [Online]. Available: http://journals.humankinetics.com/doi/10.1123/japa.7.2.129

[15] J. I. Sheikh and J. A. Yesavage, "Geriatric Depression Scale (GDS) Recent Evidence and Development of a Shorter Version." *Clinical Gerontologist*, vol. 5, no. 1-2, pp. 165–173, nov 1986. [Online]. Available: https://www.tandfonline.com/doi/full/10.1300/J018v05n01_09

[16] 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," *Journal of Aging and Physical Activity*, vol. 7, no. 4, pp. 339–353, oct 1999. [Online]. Available: http://journals.humankinetics.com/doi/10.1123/japa.7.4.339

[17] 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. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0966636215009121

[18] M. P. Hobart, R. Goldberg, J. J. Bartko, and J. M. Gold, "Repeatable battery for the assessment of neuropsychological status as a screening test in schizophrenia, II: convergent/discriminant validity and diagnostic group comparisons." *The American journal of psychiatry*, vol. 156, no. 12, pp. 1951–7, dec 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10588410

[19] 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. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/09658210244000027

[20] 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," *Journal of the International Neuropsychological Society*, vol. 7, no. 5, p. S1355617701755075, jul 2001. [Online]. Available: http://www.journals.cambridge.org/abstract_S1355617701755075

[21] 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," *Journal of Psychopathology and Behavioral Assessment*, vol. 31, no. 4, pp. 412–416, dec 2009. [Online]. Available: http://link.springer.com/10.1007/s10862-008-9123-z

[22] P. Dawadi, D. J. Cook, and M. Schmitter-Edgecombe, "Automated clinical assessment from smart home-based behavior data," *IEEE Journal of Biomedical and Health Informatics*, vol. 99164, pp. 1–12, 2015.

[23] J. W. Ratcliff and D. E. Metzener, *Pattern-matching-the gestalt approach*. MILLER FREEMAN, INC 411 BOREL AVE, SAN MATEO, CA 94402-3522, 1988, vol. 13, no. 7.

[24] 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. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0005789486800600

[25] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving Prediction of the Minority Class in

Boosting," 2003, no. December, pp. 107–119. [Online]. Available: http://link.springer.com/10.1007/978-3-540-39804-2_12

[26] 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. [Online]. Available: http://dl.acm.org/citation.cfm?id=1622407.1622416