Multimodal Fusion of Smart Home and Text-based Behavior Markers for Clinical Assessment Prediction

GINA SPRINT*, Department of Computer Science, Gonzaga University, sprint@gonzaga.edu

DIANE J. COOK, School of Electrical Engineering and Computer Science, Washington State University, djcook@wsu.edu

MAUREEN SCHMITTER-EDGECOMBE, Department of Psychology, Washington State University, schmitter-e@wsu.edu

LAWRENCE B. HOLDER, School of Electrical Engineering and Computer Science, Washington State University, holder@wsu.edu

New modes of technology are offering unprecedented opportunities to unobtrusively collect data about people's behavior. While there are many use cases for such information, we explore its utility for predicting multiple clinical assessment scores. Because clinical assessments are typically used as screening tools for impairment and disease, such as mild cognitive impairment (MCI), automatically mapping behavioral data to assessment scores can help detect changes in health and behavior across time. In this paper, we aim to extract behavior markers from two modalities, a smart home environment and a custom digital memory notebook app, for mapping to ten clinical assessments that are relevant for monitoring MCI onset and changes in cognitive health. Smart home-based behavior markers reflect hourly, daily, and weekly activity patterns, while app-based behavior markers reflect app usage as well as writing content and style derived from free-form journal entries. We describe machine learning techniques for fusing these multimodal behavior markers and utilizing joint prediction to improve the performance of automated assessment. Joint predicting a variable involves augmenting feature vectors with "joint features," which are predictions of other, related variables to improve prediction accuracy. We evaluate our approach using three regression algorithms and data from 14 participants with MCI living in a smart home environment. Using these multimodal fusion and joint prediction techniques, we observed moderate to large correlations between predicted and ground-truth assessment scores, ranging from $r = 0.601$ to $r = 0.871$ for each clinical assessment.

• Human-centered computing~Ubiquitous and mobile computing • Applied computing~Life and medical sciences~Health informatics • Computing methodologies~Machine learning

T[†]his work is supported by the National Institutes of Health, under grant R01EB009675, by the National Science Foundation, under grant 1954372, and by the Office of the Assistant Secretary of Defense for Health Affairs through the Peer Reviewed Alzheimer's Research Program, Quality of Life Research Award, Funding Opportunity Number: W81XWH-15-PRARP-QUAL under Award No. AZ150096. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense.
Authors' address: G. Sprint, Department of Computer Science, Gonzaga University, 502 E Boone Ave, Spokane, WA, 99258 USA; email: sprint@gonzaga.edu. D. Cook, and L. Holder, School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, 99164-2752, USA; emails: djcook@wsu.edu, and holder@wsu.edu. M. Schmitter-Edgecombe, Department of Psychology, Washington State University, Pullman WA 99164-4820; email: schmitter-e@wsu.edu,

Additional Keywords and Phrases: Behavior markers, Clinical assessment prediction, Natural language processing, Smart homes

ACM Reference Format: TBD

1 INTRODUCTION

Individuals with amnestic mild cognitive impairment (MCI) and Alzheimer's Disease typically experience symptoms such as memory loss, difficulty with language and visual-spatial abilities, decreased ability to focus, and issues with reasoning, planning, and complex decision making. Of these symptoms, difficulty with everyday memory exhibits the greatest degree of impairment and can negatively impact one's independence and quality of life [1]. To assist with memory impairment in performing daily activities, compensatory devices, such as pagers and memory notebooks, have been successfully introduced [2]. Recently, technology has enabled enhanced digital versions of such compensatory devices, utilizing mobile apps and smart environments. For example, a digital memory notebook app can use notifications to remind users to log important events, tasks, and notes. Such an app can also be coupled with a smart home. The smart home continuously and unobtrusively collects naturalistic data, which can be automatically labeled with corresponding activity labels such as cook, work, and sleep. Using labeled activities to provide context, smart home algorithms can automatically provide in-the-moment digital memory notebook prompts to remind residents to perform common, day-to-day activities and encourage digital memory notebook use [2], [3].

Beyond providing compensatory assistance, smart homes and memory notebook apps serve a dual purpose for MCI and Alzheimer's Disease stakeholders. Automated algorithms can continuously analyze smart home sensor data to model activity and behavior patterns over time [4]. If there is a sudden or slow onset of behavior change detected, care providers and family members can be notified and provide early treatment. Recently, researchers are using behavior markers and machine learning to map smart home data onto clinical health scores [5]–[8]. Such mappings could then be used to regularly screen for changes in health, complementing more traditional clinical assessment methods and leading to proactive intervention. These mappings are also based on continuous data, which may provide additional insights and could further augment data collected in a brief visit with a physician. When combined with additional information sources, such as demographic information or wearable data, the accuracy of these health score predictions can be further improved [5]. If an intervention tool such as a memory notebook app is introduced into a daily routine, use of the tool itself can also provide a unique source of information that may correlate with cognitive health. This process of merging data from different sources (i.e., modes or modalities) to feed machine learning problems is called multimodal fusion [9].

In this paper, we explore multimodal fusion to predict ten different clinical assessment scores using regression techniques (see Section 3.2 for clinical assessment details):

Objective Testing Scores

1. Repeatable Battery of Neuropsychological Status (RBANS)
2. Wechsler Test of Adult Reading (WTAR)
3. Delis-Kaplan Executive Function System F-A-S (FAS) test
4. Delis-Kaplan Executive Function System Design Fluency Test (DFT)

Self-report Measures

5. Prospective and Retrospective Memory Questionnaire (PRMQ)

6. Instrumental Activities of Daily Living - Compensation Scale (IADL-C)
7. Geriatric Depression Scale Short Form (GDS)
8. Quality of Life Scale – Alzheimer's Disease (QoL-AD)
9. Coping Self Efficacy Scale (CSE)
10. Satisfaction with Life Scale (SWLS)

Our data modalities include behavior markers extracted from ambient sensors embedded in smart homes, a memory notebook tablet app called EMMA (Electronic Memory and Management Aid) [10], and participant demographic information. We hypothesize that smart homes and digital memory notebooks offer informative behavior markers, that when the markers used in combination, can predict multiple clinical health scores. We validate our methods and provide evidence to support our hypothesis using data collected from N = 14 participants with amnestic MCI who participated in the EMMA/smart home partnered condition of a pilot randomized controlled clinical trial [10].

This paper offers both clinical and technical contributions. In terms of clinical contributions, we introduce and evaluate the ability to automatically assess health in naturalistic, unscripted settings. We further describe how an intervention app can provide dual use as an assessment tool as well as a compensatory device. Based on our participant sample, we offer insights on the relationship between cognitive health assessments and behavior marker sets describing activity patterns and memory notebook usage. In terms of technical contributions, we introduce methods for extracting digital markers that reflect patterns in behavior, writing content/style, and intervention adherence. We describe and compare multiple approaches to fusing these multimodal data. Furthermore, we consider the design of machine learning techniques to predict precise clinical scores. Finally, we utilize joint prediction to boost assessment performance by harnessing the predictive relationship between multiple assessment measures.

2 RELATED WORK

A growing body of research has explored mapping sensor-based features to clinical assessment scores using machine learning techniques. Studies utilizing smart home-based features and text-based features are the most relevant to the present study.

2.1 Smart Home-based Assessment Prediction

Data collected from ambient sensors installed in environments, such as homes, offices, and cities, can be used to quantify and track the activities and behaviors of people over time. Recent research has shown that features extracted from smart home data can be used to identify current or past activities [11] and forecast occurrences of future activities [12]. These methods build a foundation for modeling patterns of activities [13], detecting changes in activity-based behavior over time [4], and identifying behavioral differences between subject groups [14]. In this paper, we build on the foundation of prior activity modeling research to extract features from activity-labeled smart home data that reflect long-term behavior patterns.

Similarly, research has indicated that smart home sensor data can be analyzed to detect target health conditions. Much of the prior work focused on designing machine learning models to predict diagnostic categories such as MCI [7], [15] and loneliness among older adults [16]. Earlier studies have also mapped smart home-based features to numeric health assessment scores, including Mini-Mental State Examination, Clinical Dementia Rating [7], Repeatable Battery of Neuropsychological Status, and Timed Up and Go test [5], [6], [8]

measures. Alberdi *et al*. [8] further employed regression algorithms to predict Arm Curl test, Digit-Cancellation test, Prospective and Retrospective Memory Questionnaire, and Geriatric Depression Scale measures. Like these studies, in this paper we use smart home data to predict ten clinical assessments. Unlike these previous works, we fuse smart home data with other information sources to improve assessment accuracy.

2.2 Text-based Assessment Prediction

Several studies have used text-based features for clinical analysis. The text originates from a variety of sources, such as paper-pencil writing [17], typed text [18]–[23], and speech transcriptions [9], [19], [24]–[30]. Text features vary from simple to complex, including statistics like word count, mean number of words per sentence, and lexicon category frequencies; classic natural language processing-based features like part-of-speech proportions/ratios, readability scores, sentiment analysis, and term frequency-inverse document frequency (TF-IDF) values; and more recent neural network-based features like word/document embeddings and topic models. These text mining methods have successfully been used to detect cases post-traumatic stress disorder [22], depression [18], [26], [31], MCI [9], [17], [24], [25], and Alzheimer's Disease [17], [19], [28]–[30]; determine correlation with life satisfaction [23], [27], [28], and analyze language differences between patients pre- and post-liver transplant [20]. Text-based features are used in these applications because research has shown that one's lexico-syntactic patterns not only change over time but change more drastically with onset of health conditions. For example, the early stages of Alzheimer's Disease typically exhibit a decreased vocabulary, simplified syntax/semantics, and increased use of empty filler words [29].

While much of the prior work in text-based clinical assessment distinguishes diagnostic categories (e.g., healthy vs MCI [9]), creation of the ADReSS benchmark dataset [32] created an opportunity for researchers to design regression methods that predict numeric Mini-Mental State Examination scores based on speech features for individuals with cognitive impairments [27], [28], [30]. Ostrand and Gunstad extracted 16 linguistic features from speech transcriptions to predict current and future Mini-Mental State Examination scores for participants with cognitive impairment [27]. The researchers concluded linguistic features were good predictors of Mini-Mental State Examination scores at the time when they were recorded and one year in the future.

Most of these prior studies used text sourced from speech transcriptions for clinical domain mappings. Fewer studies have used health-related text originally sourced by study participants in a written form, like we are proposing in this paper. Dreisbach *et al.* review methods for mining patient-authored text for symptom extraction [33]. Other studies that analyze written text include one analyzing text messages from patients to care providers [20] and one analyzing self-narratives describing traumatic experiences [22]. Dickerson *et al*. [20] analyzed patient-authored messages to care providers from controls and patients with end-stage liver disease pre- and post-liver transplant. The researchers used 19 natural language processing features extracted from the messages to detect differences in language between controls and patients pre- and post-transplant. He *et al*. [22] similarly extracted natural language processing-based features from online surveys that included free-response questions about traumatic experiences and related symptoms. The responses were pre-processed into unigrams, bigrams, and trigrams, then mined for keywords related to stress. Combining free-form text features with multiple choice survey responses, the researchers trained classifiers to predict the presence or absence of post-traumatic stress disorder. The authors reported accuracy improvements from 0.94 (no text-features) to 0.97 (using fusion with the text features).

The work we present in this paper builds on these prior studies. As with the approaches that predict Mini-Mental State Examination scores, we introduce natural language processing-based methods to predict numeric clinical scores. Unlike the prior work, we mine a core set of text features that are used to predict multiple clinical measures and further utilize the relationship between the measures to boost predictive performance. Additionally, we fuse natural language processing with mining of smart home sensor data and intervention app usage data to offer a more comprehensive assessment of a person's health status.

2.3 Multimodal Fusion

New forms of technology have enabled data collection from many different sources, or modes. Often data from disparate sources, such as motion sensors and audio, are combined as input to machine learning algorithms to solve health-related problems and/or make predictions. This fusion process can improve machine learning results because the different modes typically capture different (though possibly redundant) aspects of the same process [9], providing a more holistic view. Researchers in a variety of health-related contexts have explored different approaches to combine data from multiple modalities. For example, researchers have combined speech and text to detect MCI [34], Alzheimer's Disease [28], and depression [26]. The most common approaches include early fusion, late fusion, and hybrid fusion [35]. Using early fusion, features from different modes are concatenated for input to a machine learning algorithm. With late fusion, features from different modes are input to separate machine learning algorithms; then a second algorithm learns how to combine the mode-specific predictions into a final target variable. While hybrid fusion approaches vary in design, one example is stacked generalization fusion, which is an ensemble approach to combining predictions from diverse modality classifiers. Recently, Alkenani *et al*. implemented stacked fusion to predict Alzheimer's Disease using speech and writing datasets [19].

In the current paper, we combine data from different modes, including demographic information, smart home sensors, digital memory notebook usage, and patient-authored text. Gosztolya *et al*. utilized linguistic and acoustic features from speech recordings to detect MCI, finding the best results were achieved using a late fusion of the two feature types [34]. However, little work focuses on fusing information from more heterogeneous sources. In one published case, Fraser *et al*. explored early and late multimodal fusion (at various levels) of comprehension questions, eye-tracking data, and audio recordings (speech and transcription) to classify 55 participants as MCI or healthy control [9]. Using natural language processing techniques, Fraser *et al*. extracted language features from the transcription text, including total words, mean sentence length, phrase type proportions (prepositional, noun, and verb groups), and part-of-speech ratios. Using logistic regression and support vector machines, the researchers concluded that the best classification results were achieved using a variation of late fusion. The work we present here represents a new direction for fusing health-related information sources. While fusing behavior data with app usage and free-form text requires combining dramatically different types of features, all of these information sources have independently been identified as indicators of cognitive health. We therefore conjecture that the fusion of these data will provide an enhanced ability to predict clinical health measures.

2.4 Joint Prediction

Recent machine learning research has utilized joint prediction as a strategy to improve prediction performance [5], [12], [36]. Typically, with joint prediction, the predictions of a target variable are combined with predictions

of other, related variables, to expand the feature vector that is fed to a machine learning algorithm. Minor *et al*. used joint prediction to improve forecasting of smart home activity occurrences using multi-output regression [12]. To do this, the researchers first performed independent prediction (called the baseline model) by utilizing sensor-based features to predict the time until each smart home activity would occur. Next, the authors designed joint prediction by adding each activity's previous occurrence predictions to the original sensor-based features as joint features. Using this joint prediction technique, the researchers were able to achieve an 85.11% decrease in prediction error compared to the baseline model. Sprint *et al*. investigated patient similarity and joint prediction to improve Functional Independence Measure motor score prediction for 27 inpatient rehabilitation participants [36]. The authors utilized features extracted from wearable sensors while participants performed activities in an ecological rehabilitation environment. For joint prediction, the sensor-based features were augmented with Functional Independence Measure score predictions for similar participants in the training set as joint features. The best leave-one-out cross validation Functional Independence Measure regression results were achieved when predictions utilized the joint features (r = 0.88).

More recently, Cook and Schmitter-Edgecombe applied joint prediction to predict seven clinical health assessment scores (Wechsler Test of Adult Reading, Telephone Interview of Cognitive Status, Repeatable Battery of Neuropsychological Status, Delis-Kaplan Executive Function System F-A-S test, Timed Up and Go, Dysexecutive questionnaire, and Alzheimer's Disease Cooperative Study Activities of Daily Living Inventory) from smartwatch and smart home sensor data [5]. For this study, clinical assessment scores, smart home data, and smartwatch data were collected from 21 older adults. Using these data, a set of digital behavior markers were extracted from each sensor source and concatenated to form a participant's "behaviorome." The behaviorome was used as input to independent and joint prediction regression models to predict each score. Some regression models were trained using the smart home and smartwatch data separately, while some were trained using a combination of the two sensor modalities. The greatest performance was achieved for each assessment using joint prediction over independent prediction. One of the seven assessments exhibited the best performance using a fusion of features from both modalities (Telephone Interview of Cognitive Status). The remaining six assessments were best predicted using either smart home data alone or smartwatch data alone. This earlier work represents the first use of joint prediction of multiple clinical measures found in the literature. In this paper, we extend the technique of joint prediction to encompass an expanded set of clinical measures. As in the previous study, we anticipate that assessment performance will be enhanced through joint prediction because of the predictive relationship that is inherent between multiple clinical measures. We further extend the earlier work by including additional, and more varied, sources of information than sensor-based behavior markers. Specifically, we fuse text markers and app use with the behavior markers for the joint prediction models.

3 METHODS

Our approach to clinical assessment prediction utilizes a custom digital memory notebook app, EMMA, with smart home integration (Section 3.1), an EMMA/smart home-based data collection protocol (Section 3.2), extracted smart home-based behavior markers (Section 3.3), extracted EMMA text-based behavior markers (Section 3.4), and clinical assessment prediction making use of the markers (Section 3.5). The following sections describe these facets of the study.

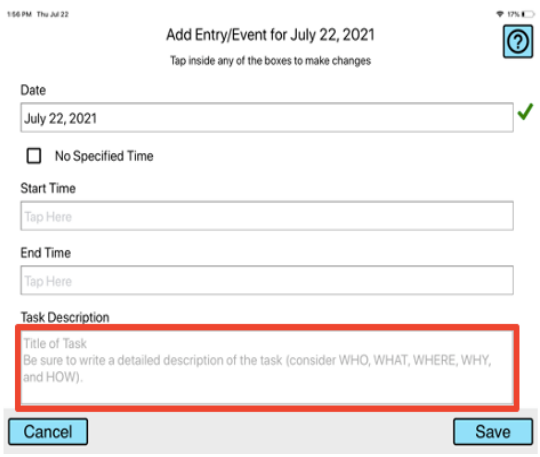
3.1 Digital Memory Notebook Application and Smart Home Integration

To help individuals compensate for memory difficulties, we introduce a digital memory notebook, called EMMA, that runs on an iPad [3], [10], [37]. EMMA's interface was designed for older adults and individuals with MCI [2]. EMMA supports two text-based functions, "tasks" and "notes". A task, highlighted in Figure 1(a), consists of a date, optional start and end times, and a description. Users create task descriptions that provide details about the task "who", "what", "where", "why", and "how", such as:

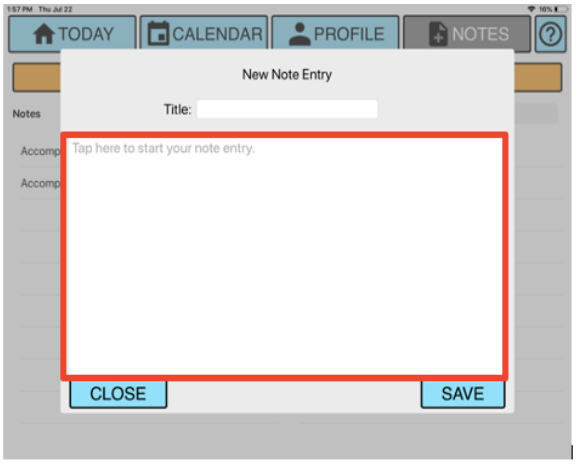
- "On [date], take [item] for [person] to [address]"
- "Fill medicine box [frequency]"
- "Schedule appliance repair for [date] at [time]"

In these examples, the brackets are placeholders for specific information an EMMA user might include. From any screen, tapping on the notes tab button at the top brings the user to the notes screen where the user can view and edit previously created notes, as well as add a new note. Figure 1(b) shows a screenshot of the notes screen. An EMMA note is free-form text that allows a user to document personal lists, processes, projects, and long-term goals. Here are a few examples:

- "At grocery store, be sure to pick up milk, eggs, bread, apples, …"
- "When filling medicine box, put 2 of [prescription] in each day's compartment…"



(a) The "add a new task" screen with the task description textbox shown in a red outline.



(b) The "add new note" screen with the note content textbox shown in a red outline.

Figure 1: EMMA app screenshots of the two screens where free-form (a) task and (b) note text are entered by the user.

Task and note content and metadata, as well as all other recorded EMMA interactions, are saved to a cloud database. These data include timestamps for each visited EMMA screen and each user tap. From these data, we extract EMMA usage features describing the time and duration of user interactions with each app screen. From the free-form task and note text we utilize natural language processing to mine informative features about an EMMA user and their behavior. We hypothesize that such EMMA usage and text features are predictive of clinical assessments, such as the ones described in Section 3.2.

A unique aspect of the EMMA notebook is its interface with the CASAS smart home technology [38]. Ambient sensors are installed throughout the smart homes which include passive infrared motion sensors, ambient light sensors, magnetic door sensors, and ambient temperature sensors. To model home-based behavior, time series data collected from the sensors are analyzed with activity recognition algorithms [11]. Activity recognition maps sensor readings (e.g., new motion near the kitchen stove, front door opens) to a pre-defined set of activity labels. We evaluate recognition accuracy in previous studies for the activities bed-toilet transition, cook, eat, enter home, leave home, personal hygiene, relax, sleep, wash dishes, and work [11].

As activities are detected, the information is sent from activity recognition through the smart home middleware to EMMA [3]. A RabbitMQ message broker sends a stream of messages in real time from the smart home to EMMA, containing recognized activity labels. These labels are sent to a Python-based event listener to store in the database. EMMA receives updated activity information by sending requests to a Flask web application. This Flask application further stores all user interactions with EMMA, which we analyze in this paper. We hypothesize that the collected smart home-based features can provide additional predictive indicators of clinical measures.

3.2 Study Design

The data used for analysis in the present study represents data from 14 individuals with amnestic MCI who participated in the EMMA/smart home partnered condition of a five-month long pilot randomized clinical trial [10]. The trial enrolled 32 participants with amnestic MCI and was designed to examine whether learning and sustained use of EMMA could be augmented through partnership with a smart home that initiated activity-aware, transition-based smart prompts to engage with EMMA. All participants met criteria for amnestic MCI [39]. These criteria included (a) self- or informant-report for 6 months or more of memory complaints, (b) objective cognitive impairment in the memory domain; generally, 1.5 standard deviations below appropriate norms taking into account premorbid abilities, (c) did not meet criteria for the Diagnostic and Statistical Manual of Mental Disorders Major Neurocognitive Disorder (DSM-5), and (d) no severe depression at start of intervention (GDS score < 10). Upon enrollment, participants were randomly assigned to either the EMMA/smart home partnered condition (N = 15) or the EMMA only condition (N = 17). For the purposes of this paper, we exclusively use the data collected from the participants in the EMMA/smart home partnered condition due to the availability of smart home data, and therefore smart home-based behavior markers. These participants had a CASAS smart home in a box (SHiB) installed in their home which served as a second longitudinal data collection modality [38] in addition to the data collected throughout the study via the EMMA app. Because one person in the partnered condition did not complete training, the data in this study represent the 14 participants who completed training and three additional months of data collection. Table 1 provides an overview of the demographics for the participants in the EMMA/smart home partnered condition included in this work.

Table 1. EMMA/smart home condition participant demographics.

N	Gender	Age (years)	Highest Grade of Education	Marital Status
14	4 male; 10 female	74.429 ± 5.653	17.07 ± 2.235	4 widowed; 6 married or domestic partner; 2 divorced; 2 single

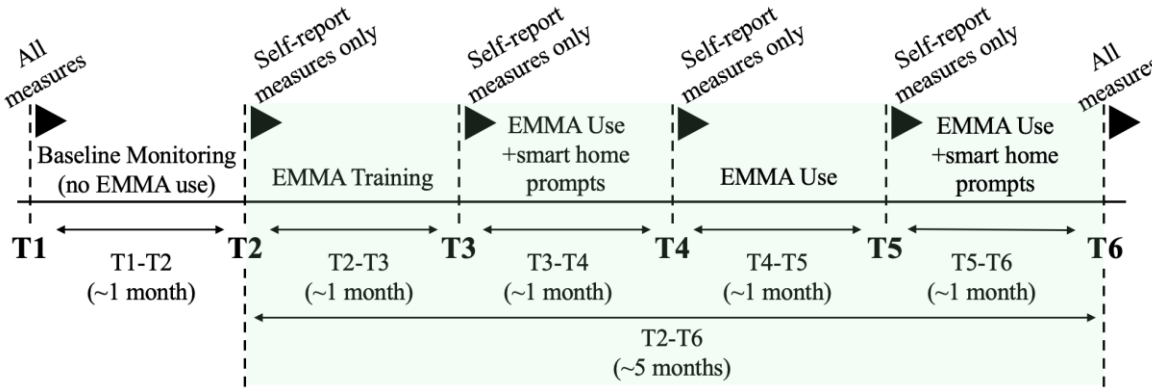


Figure 2: Study timeline. EMMA=Electronic Memory and Management Aid, T=timepoint. Note: All clinical data used in this study was collected at timepoint T6.

Data collection for the EMMA study participants was divided into approximately five one-month periods. The periods are separated by six timepoints (T1 through T6) when four objective and six self-report clinical assessments were administered, and portions of an EMMA intervention were delivered (the intervention is described in detail below). Figure 2 shows a timeline of the timepoints that segment the total data collection duration into five distinct periods (T1-T2, T2-T3, T3-T4, T4-T5, and T5-T6). At T1 and T6, four objective laboratory assessments were administered, including RBANS [40], the Wechsler Test of Adult Reading (WTAR) [41], the Delis-Kaplan Executive Function System F-A-S (FAS) test, and the Delis-Kaplan Executive Function System Design Fluency Test (DFT) [42]. RBANS is a reliable and well-validated suite of 12 subtests that measure cognitive abilities related to attention, language, visuospatial/constructional abilities, and memory. WTAR measures verbal knowledge and intellectual skills via a reading test. Subjects are given a list of 50 words with irregular pronunciation and are instructed to read the words aloud. The FAS assesses executive functioning by having subjects state as many words as possible that start with the letters F, A, or S in a one-minute time window. Lastly, the DFT assesses executive function, and more specifically problem-solving and visual processing, by asking subjects to complete designs by connecting dots. Age-corrected standard scores were used in the analyses.

In addition to these objective laboratory assessments, several self-report questionnaires were administered at each timepoint (T1, …, T6) throughout data collection. Raw scores were used for the questionnaire data. These included the Prospective and Retrospective Memory Questionnaire (PRMQ) [43], Instrumental Activities of Daily Living - Compensation Scale (IADL-C) [44], the Geriatric Depression Scale Short Form (GDS) [45], the Quality of Life Scale – Alzheimer's Disease (QoL-AD) [46], the Coping Self-Efficacy Scale (CSE) [47], and the Satisfaction with Life Scale (SWLS) [48] assessments. The PRMQ is a 16-item questionnaire measuring various types of prospective and retrospective memory, including short-term, long-term, self-cued, and environmentally cued memory. The IADL-C is a 27-item scale that assesses early functional difficulties and compensatory strategies. It has four domain subscales, including money/self-management, daily living, travel/event memory, and social skills [44]. The shorter version of the GDS is a 15-item questionnaire detecting depression in the elderly, physically ill, and mildly demented. The QoL-AD is a 16-item scale assessing quality of life via five

different domains, including material/physical well-being, relationships, social/community/civic activities, personal development/fulfillment, and recreation. The CSE is a 26-item assessment that measures a subject's perceived confidence and ability to cope with various life challenges. Lastly, the SWLS asks subjects five Likert-scale questions about satisfaction with their life in its entirety (as opposed to specific domains of their life, such as finance). Table 2 shows the mean and coefficient of variation (CV) for each clinical measure at timepoint T6 for the 14 participants. These ten T6 scores are predicted for each clinical assessment (RBANS, WTAR, FAS, DFT, PRMQ, IADL-C, GDS, QoL-AD, CSE, and SWLS) for each participant using behavior marker data collected throughout the study (see Sections 3.3 and 3.4 for behavior marker details; see Section 3.5 for clinical assessment prediction details).

Table 2: Clinical scores and their distributions for EMMA/smart home condition participants at timepoint T6.

	Measure	**Construct Assessed**	**(Min, Max)**	**Mean ± CV%**
Objective Laboratory Assessments	Repeatable Battery of Neuropsychological Status (RBANS)	General neurocognitive status	(66, 110)	92.214 ± 13.980%
	Wechsler Test of Adult Reading (WTAR)	Verbal intellectual abilities	(84, 123)	112.857 ± 10.292%
	Verbal Fluency on F-A-S letter fluency (FAS)	Executive function	(5, 19)	12.214 ± 30.155%
	Design Fluency Test (DFT)	Executive function	(8, 18)	11.214 ± 22.983%
Self-report Questionnaires	Prospective and Retrospective Memory Questionnaire (PRMQ)	Memory	(30, 62)	42.143 ± 23.265%
	Instrumental Activities of Daily Living - Compensation (IADL-C) scale	Everyday functioning and compensatory strategies	(1.370, 4.111)	2.414 ± 35.385%
	Geriatric Depression Scale (GDS)	Depression	(0, 11)	3.184 ± 95.260%
	Quality of Life Scale (QoL-AD)	Quality of life	(26, 49)	38.500 ± 18.889%
	Coping Self Efficacy (CSE)	Coping abilities	(66, 128)	95.500 ± 19.228%
	Satisfaction with Life Scale (SWLS)	Satisfaction with life as a whole	(11, 33)	25.214 ± 25.899%

The duration from T1-T2 served as a baseline data collection period in which smart home data were collected for activity recognition training purposes. Participant use of EMMA did not begin until T2, which was the start of the EMMA training intervention. The intervention included 5-6 formal training sessions with a clinician on how to use EMMA during the T2-T3 period. Details about the intervention, which were conducted in participants' homes, can be found in prior work [10], [37]. During post-intervention months T3-T6, participants were encouraged to continue to use EMMA to record information about past activities, to help them remember to do important tasks and to support day-to-day activities, including both short- and long-term goals. Furthermore,

during T2-T3, T3-T4 and T5-T6, smart home activity recognition-aware prompts were delivered to the study participants via EMMA to remind them to use EMMA. Smart home reminders to use EMMA were not delivered to participants during T4-T5.

From the collected participant information and two data collection modes, the smart home and the EMMA app, we extract five different sets of behavior markers to model an EMMA study participant's behavior:

1. DEMO: T1 Demographics (age, gender, education level, and marital status; see Table 1)
2. SHBM: T2-T6 Smart home behavior markers (see Section 3.3)
3. USAGE: T2-T6 EMMA app usage behavior markers (see Section 3.4)
4. NOTE: T2-T6 EMMA app note entry text-based behavior markers (see Section 3.4)
5. TASK: T2-T6 EMMA app task description text-based behavior markers (see Section 3.4)

As a pre-processing step, any feature with near-zero variability across the participant periods was removed and not used for prediction. For the demographic features, this included the ethnicity and race features which exhibited zero variability across features (all participants were non-Hispanic/Latino and Caucasian).

3.3 Smart Home Behavior Markers (SHBM)

In recent work, we created a set of digital behavior markers for a smart home resident by extracting 556 features from activity-labeled smart home data [5]. To do this, we computed features that describe behavior at the hour level, then aggregated these features at the day level. Additional features describing a person's overall routine over several days were then computed by aggregating the daily features over a longitudinal period. Examples of each of these time period-based behavior markers are as follows:

- Hourly: # of sensor readings, # of distinct activities performed, # of distinct locations visited, time spent on each activity, time spent at each location.
- Daily: Daily totals for each of the hourly features, time of day for first occurrence of each activity, time of day for first visit to each home location.
- Overall: Statistics for hourly and daily behavior markers (e.g., mean, median, standard deviation, max, min), regularity indices, circadian rhythm strength.

Complete descriptions of the extracted smart home behavior markers are detailed in the literature [5][1]. For the current study, we extract these 556 behavior markers from the smart home data collected between periods T2-T6 to form the smart home behavior marker set.

3.4 Digital Note Behavior Markers (USAGE, NOTE, TASK)

To expand upon our prior work with smart-home based behavior markers, we explore new behavior markers computed from a different mode of data collection, the EMMA app. Table 3 summarizes these EMMA-based behavior markers that we extract between data collection periods T2-T6[1]. First, from the EMMA app database, we extract USAGE behavior markers. These are similar to the smart home behavior markers (e.g., # of distinct uses, time spent in the various screens of the app, time of day for first use, regularity indexes, circadian rhythm strength), but based on the participant interactions with EMMA. In addition to quantifying a participant's usage of the EMMA app, we are interested in utilizing information from the tasks and notes a participant enters in EMMA. This text contains valuable information about a person's daily and overall behavior routine and goals,

[1] Code to extract smart home and text digital behavior markers is available at https://github.com/WSU-CASAS/DM

as well as a person's writing style. To tap into this information, we first extract two different sets of free-form participant text from the EMMA database, the task description text (TASK), and the note content text (NOTE).

Prior to computing behavior markers from these two categories of text, we perform pre-processing to clean the text and prepare it for different types of natural language processing. First, we tokenize the text into phrases and extract only the unique phrases authored by a participant during each period. We do this because notes can be edited and tasks can be scheduled to repeat, causing some phrases to appear multiple times within a period. With the unique phrase text, we extract initial features for which text artifacts like capitalization, punctuation, and verb tense are important. These include named entity recognition features (e.g., the number of specific datetimes, locations, organizations, and persons) and Linguistic Inquiry and Word Count (LIWC) features [49]. LIWC is a standard tool used in computational linguistics to extract features that are relevant for psychological analyses. LIWC is primarily comprised of a hierarchical dictionary that maps over 86% of common words used in writing and speech to categories. The categories include 21 linguistic dimensions (e.g., percentage of words that are different parts of speech), 41 psychological construct categories (e.g., percentage of words that are related to affect, cognition), 6 personal concerns categories (e.g., percentage of words that are related to work, home), and 5 informal language categories (e.g., percentage of words that are related to assents, fillers). In addition to these category-based features, LIWC produces additional features such as total word count, words > 6 letters, summary language composite scores (e.g., percentiles measuring analytical thinking, clout, authentic, and emotional tone measures), and punctuation use (e.g., percentage of text that are periods, commas).

Next in the pre-processing pipeline, we follow standard text normalization steps to prepare for additional natural language processing techniques [21], [22], [50], including:

- Converting text to lowercase
- Replacing contractions with their expanded form
- Removing punctuation marks (e.g., replace "&" with "and")
- Removing numeric characters
- Removing stop words (using Python NLTK stop words)
- Performing word stemming (using Python NLTK Porter Stemmer)

We then compute natural language processing features that have exhibited predictive power in previous, related work [18], [28], [29], including readability scores from the Python Textatistic library (e.g., Dale-Chall score, SMOG score), sentiment analysis features using the Python TextBlob library (e.g., polarity, subjectivity), term frequency-inverse document frequency (TF-IDF) values for the 100 most frequent unigrams and bigrams using Python's Sci-Kit Learn library, word embeddings using the Python Gensim library, and Latent Dirichlet Allocation topic modeling with 5 latent topics using Sci-Kit Learn (see Table 3 for a list of all EMMA-based behavior markers). To elaborate, we compute TF-IDF features by fitting a TF-IDF vectorizer on the training text and using this vectorizer to transform the testing text. From these TF-IDF vectors, only the values for the 100 most frequent unigrams and bigrams in the training text are included as features. For word embeddings, we explored several Word2Vec models, including pre-trained models and a model trained on the available training set text each fold of cross validation. The best results were obtained using a Word2Vec model named glove-wiki-gigaword-50, which is pre-trained on the Wikipedia 2014 + Gigaword 5 dataset to obtain a 50-dimensional vector for each word. Following Gonzalez-Atienza *et al*., all the word vectors in a period are averaged to produce

a 50-dimensional vector summarizing the dominant word embeddings for the period. From this vector two additional features, the mean and standard deviation, are extracted [29].

Two of these natural language processing feature types, TF-IDF and topic modeling, require information extracted over a corpus and are therefore computed right before training on the pre-processed training text from each fold of cross validation. We do this to prevent data leakage from the test set to the training set when using cross validation. As an example of how data leakage can occur, consider computing document frequencies over the entire corpus (all participant text) and then using these frequencies to compute TF-IDF vectors for each participant's feature vector. In this manner, artifacts from feature vectors held out for testing are embedded in training feature vectors' data (e.g., the document frequencies).

Table 3: EMMA-based behavior markers.

Behavior Marker Set	Feature	# Features	Reference
USAGE (21 markers)	Mean daily # of taps in the app	1	
	Mean daily # of distinct uses (5 mins of inactivity has passed between uses)	1	
	Mean daily total minutes used	1	
	Mean daily first use (minutes past midnight)	1	
	Mean daily minutes spent on app screens (login, main, help, event, note/journal, milestone, reminder)	7	
	Regularity index (within weeks, within days, within each day of week)	9	[51]
	24-hour circadian rhythm (measured over the period)	1	[5]
NOTE and TASK (265 markers each)	Text structure (% unique phrases, % stop words, % unknown words, mean # of characters per word, mean # of syllables per word)	5	
	Readability (Dale-Chall score, SMOG score)	2	[28]
	# of named entities (datetimes, locations, organizations, persons)	4	
	Sentiment analysis (mean/std sentence polarity and subjectivity)	4	
	LIWC features (summary dimensions, affect, social, cognitive processes, perceptual processes, biological processes, drives, time orientation, relativity, personal concerns)	93	[18]
	TF-IDF features*	100	[18], [28]
	Word embeddings (50-length vector, mean/std of vector)	52	[28], [29]
	Latent Dirichlet Allocation topic modeling (probabilities for 5 latent topics)*	5	[18]

* = Computed from the training set of each fold of cross validation.

3.5 Clinical Assessment Prediction

We hypothesize that smart home and EMMA-based digital markers provide information about a person's overall behavior routine and therefore may be predictive of various clinical measures of behavior. To investigate this hypothesis, we conduct several machine learning experiments to predict each of the assessments described in Table 2 at timepoint T6 using behavior markers from T2-T6. Our experiments explore different applications of recent developments in joint prediction and multimodal fusion.

3.5.1 Joint Prediction

Joint prediction is a branch of machine learning research that aims to leverage predictions as "joint" features. What types of predictions are used as joint features are problem-specific; however, they are typically predictions in a similar feature space as the target variable *Y*. For example, recent research has explored the following possibilities for joint features:

- Oracle features: ground truth *Y* values for similar and/or complementary training instances [12], [36].
- Predicted features: predicted *Y* values ($\hat{Y}$) for similar and/or complementary training instances [12], [36].
- Related variable prediction features: predicted values for variables similar and/or complementary to target variable *Y* for a testing instance [5].

In this paper, we explore utilizing related variable predictions as joint features. Using joint prediction, a separate model for each predicted clinical measure is trained using behavior markers. To improve prediction accuracy, a second set of models is then trained to predict each measure using the original features combined with the predicted scores of all other measures. In this manner, prediction for a new participant's clinical assessment *A* at timepoint T6 is a set of *n* values for assessments $A \in$ {RBANS, WTAR, FAS, DFT, PRMQ, IADL-C, GDS, QoL-AD, CSE, SWLS}. The predicted values are based on predictions for the $n - 1$ assessments in the set, without needing to administer the assessments. We hypothesize that predictions of measures assessing different clinical constructs are informative because of the predictive relationships that exist between the constructs themselves. We design our experiments to include two main components, independent predictions and joint predictions. Figure 3 provides an overview of this two-part prediction process. First, for each assessment *A*, an independent predictor utilizes an input feature matrix comprised of some subset of the five behavior marker sets listed in Section 3.2. Using cross validation, the independent predictor produces a single clinical assessment score prediction for each participant, producing a vector of predicted scores. All independent assessment predictions are combined into an independent prediction matrix where each column is an assessment vector and each row represents a participant.

During joint prediction, this independent prediction matrix is joined with the original feature matrix to produce a new input feature matrix. For each assessment, the independent prediction for this assessment is removed from the joint feature matrix so each joint predictor only has access to independent features and predictions from other assessments. This step prevents data leakage and ensures the machine learning algorithms explore the predictive power of the independent and joint features. In the same fashion as the independent prediction vectors, the joint prediction vectors are combined to produce a final joint prediction matrix to be used for evaluation. Joint prediction can offer performance improvements when the target prediction variables are related, as is the case for our experiments. We hypothesize that when predicting a clinical assessment score at timepoint T6, including joint features (e.g., same-subject predictions for other complementary clinical assessments measured at timepoint T6) will improve prediction accuracy because of joint features capture different aspects of behavior.

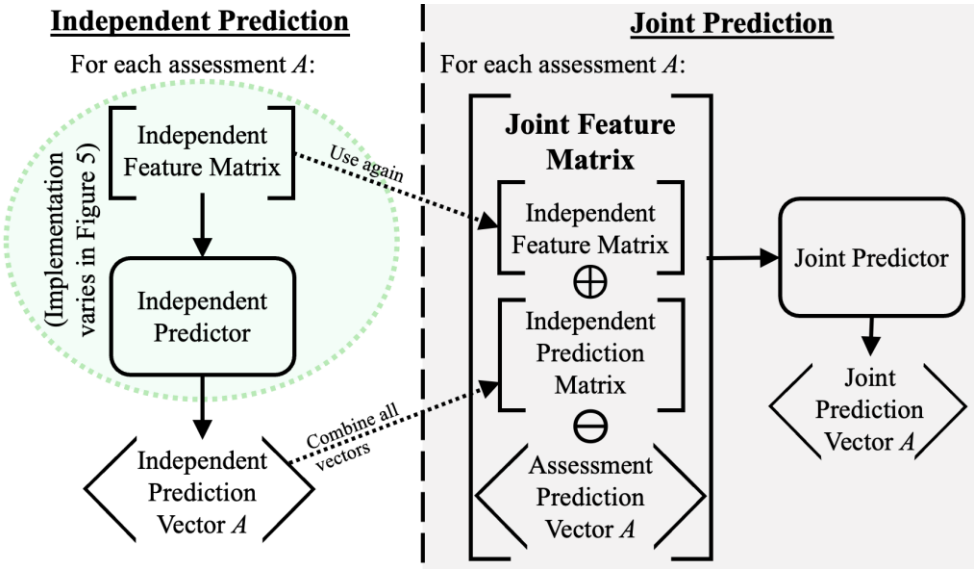


Figure 3. Independent (white background) and joint (gray background) prediction workflows for predicting participant scores on a clinical assessment. Steps needed to prepare for joint prediction are shown with dashed arrows. Highlighted in the green circle with the dashed outline are the parts of Figure 3 that vary in implementation across Figure 4(a)-(c).

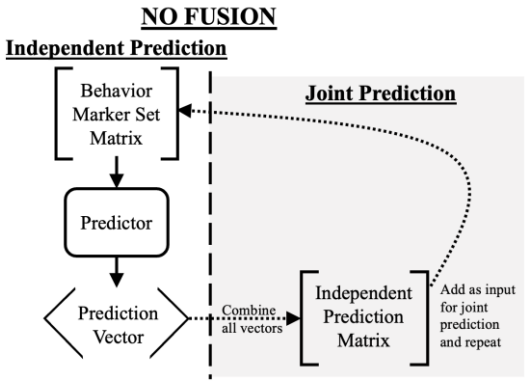
3.5.2 Multimodal Fusion

Recent work by Fraser *et al*. explored using two different variations of multimodal fusion to classify participants as MCI or not MCI [9]. The variations were early fusion, where features from different modes are combined before serving as input to a machine learning algorithm, and late fusion, where each mode's features are input to their own machine learning algorithm. With late fusion, there is an additional machine learning algorithm that combines the predictions from the individual modality's machine learning algorithm. Benefits to early fusion include the ability to model relationships between features extracted from different modes and less computational complexity since only one algorithm needs to be trained. A drawback of early fusion is the resulting high-dimensional feature space that accompanies combining features from different modes. Late fusion addresses this disadvantage by keeping the dimensionality of each algorithm's input small. Another benefit of late fusion includes the ability to use different machine learning algorithms for each mode, allowing for more fine-grained tuning [52]. This flexibility is a trade-off of increased computational complexity for late fusion.

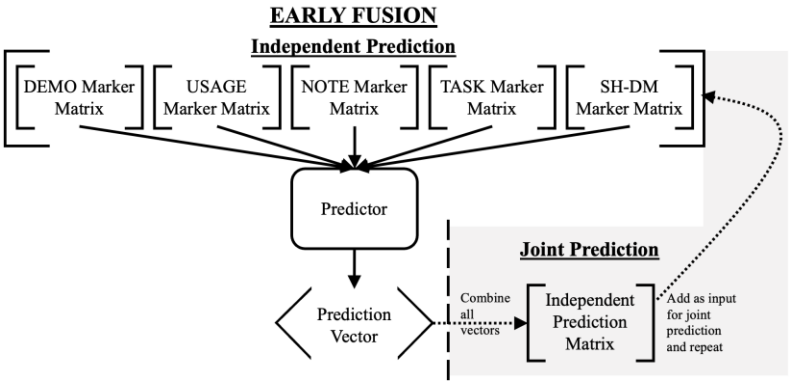
In the current paper, we evaluate the predictive impact of using early and late fusion in conjunction with joint prediction, aiming to achieve the best possible clinical assessment prediction results. For each experiment, we compare different approaches to fusing behavior marker sets. This is to explore the feature space and determine the most predictive behavior markers for each assessment. The three main types of experiments we conduct are as follows:

1. NO FUSION: independent and joint prediction for each individual behavior marker set, one at a time (see Figure 4(a) for a diagram)

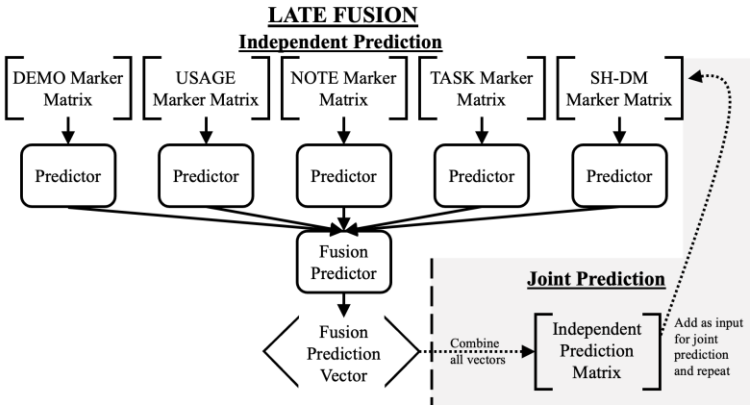
2. EARLY FUSION: independent and joint prediction for all possible behavior marker set combinations (see Figure 4(b) for a diagram)
3. LATE FUSION: independent and joint prediction with one predictor for each behavior marker set (see Figure 4(c) for a diagram)



(a). NO FUSION experiment design.



(b). EARLY FUSION experiment design.



(c). LATE FUSION experiment design.

Figure 4. Multimodal (a) NO FUSION, (b) EARLY FUSION, and (c) LATE FUSION for independent prediction of participant scores on a clinical assessment. Additional steps needed for joint prediction are shown in gray boxes with dashed arrows.

3.5.3 Feature Selection and Reduction

Because the dimensionality of the feature space is quite high for EARLY FUSION, we explore feature selection and feature reduction techniques to further improve EARLY FUSION prediction results. In the case of feature selection, we utilize recursive feature elimination with cross validation to find the optimal size and set of features for prediction [53]. In the case of feature reduction techniques, we implement principal component analysis with the number of components set to the number of instances. We additionally explore using a neural network-based autoencoder with various values for the number of bottleneck features (e.g., the resulting compressed feature vector size, *m*) [54]. An autoencoder is composed of an encoder and a decoder. The encoder is used to compress input data into a latent-space representation, which is typically a lower dimensionality (*m*) than the input dataspace. The decoder reconstructs the data from its latent-space representation back into its original form. Once trained and validated, the autoencoder's encoder can be used in isolation from the decoder as a feature reducer. For EARLY FUSION feature reduction, each fold of cross validation we train an autoencoder on the fold's training feature matrix, holding one training instance's feature vector out to use for autoencoder validation. Because the autoencoder is reducing the training feature space only, the fold's ground truth *Y* values (i.e., clinical assessment scores) and testing data are not provided to the autoencoder, ensuring data leakage does not occur. After validation, the encoder part of each fold's autoencoder is then used to reduce the feature set before a regression algorithm is trained.

3.5.4 Prediction Experiment Evaluation

To evaluate our experimental design approaches with joint prediction and multimodal fusion, resulting independent and joint prediction matrices are evaluated separately using standard machine learning performance evaluation metrics. Since the predictions are numeric, the following regression evaluation metrics are applied:

- Correlation coefficient *r* and associated *p*-value (α = 0.01; Bonferroni corrected α = 0.01 / 8 clinical assessments = 0.00125; conservatively rounded down to 0.001)
- Root mean squared error:

$$Root\ mean\ squared\ error = \sqrt{\frac{\sum_{i=1}^{n}(Y_{actual,i} - Y_{predicted,i})^2}{n}}$$

- Min-max normalized root mean squared error:

$$Normalized\ root\ mean\ squared\ error = \frac{root\ mean\ squared\ error}{(Y_{actual,max} - Y_{actual,min})} \times 100\%$$

All experiments are conducted using leave-one-out cross validation and comparison of three machine learning algorithms: decision tree regressors, random forest regressors with 100 estimators, and gradient-boosted regressors with 100 estimators. We use decision tree-based ensemble approaches because of their

demonstrated success with previous studies that, like the current study, are also characterized by a large number of features [5], [55].

4 RESULTS

We analyze unique text authored by participants in the EMMA app that totaled 46,754 words in their EMMA notes and 41,429 words in their EMMA task descriptions. These words are used to extract natural language processing-based NOTE and TASK behavior markers that we use with demographics markers (DEMO), smart home markers (SHBM), and EMMA usage markers (USAGE). We evaluate how our behavior markers can be used for clinical assessment prediction by conducting several experiments using leave-one-out cross validation. Table 4 summarizes the numerically highest NO FUSION correlation results for each assessment using independent prediction and joint prediction.

Table 4: NO FUSION results: Numerically highest correlations shown for independent (I) and joint (J) prediction experiment configurations. Bold correlation denotes numerically higher correlation result between I and J rows.

Assessment		Marker Set	Algorithm	*r*
RBANS	I	TASK	GBR	0.738*
	J	**TASK**	**GBR**	**0.810**†**
WTAR	I	USAGE	GBR	0.497
	J	**USAGE**	**DTR**	**0.502**
FAS	I	DEMO	DTR	0.393
	J	**TASK**	**DTR**	**0.537**
DFT	I	NOTE	GBR	0.407
	J	**NOTE**	**DTR**	**0.627**
PRMQ	I	**TASK**	**GBR**	**0.506**
	J	TASK	GBR	0.465
IADL-C	I	**SHBM**	**DTR**	**0.264**
	J	NOTE	DTR	0.113
GDS	I	**TASK**	**DTR**	**0.683***
	J	SHBM	RFR	0.349
QoL-AD	I	**SHBM**	**DTR**	**0.574**
	J	SHBM	RFR	0.362
CSE	I	TASK	GBR	0.506
	J	**TASK**	**GBR**	**0.577**
SWLS	I	SHBM	GBR	0.403
	J	**SHBM**	**GBR**	**0.508**

* = $p < 0.01$, ** = $p < 0.001$, † = overall assessment experiment configuration across Tables 4-7 with the numerically highest correlation, DTR = decision tree regressor, GBR = gradient-boosted regressor, RFR = random forest regressor.

To determine if improved prediction accuracy could be achieved using EARLY FUSION, we perform leave-one-out cross validation with all possible feature set combinations fused prior to training. The highest performing EARLY FUSION correlation results for independent and joint prediction are shown in Table 5.

Table 5: EARLY FUSION results: Numerically highest correlations shown for independent (I) and joint (J) prediction experiment configurations. Bold correlation denotes numerically higher correlation result between I and J rows.

Assessment		Marker Sets	Algorithm	*r*
RBANS	I	TASK, NOTE, USAGE	GBR	0.758*
	J	**TASK**	**GBR**	**0.810**†**
WTAR	I	**DEMO, USAGE**	**DTR**	**0.563**
	J	USAGE	DTR	0.502
FAS	I	DEMO, TASK	DTR	0.401
	J	**TASK, USAGE**	**DTR**	**0.557**
DFT	I	TASK, SHBM, USAGE	DTR	0.721*
	J	**TASK, SHBM**	**DTR**	**0.737***
PRMQ	I	**DEMO, TASK, USAGE**	**GBR**	**0.588**
	J	DEMO, TASK, USAGE	GBR	0.524
IADL-C	I	**DEMO, SHBM, USAGE**	**DTR**	**0.345**
	J	NOTE, SHBM, USAGE	DTR	0.251
GDS	I	SHBM, USAGE	DTR	0.853**
	J	**DEMO, SHBM, USAGE**	**DTR**	**0.861**†**
QoL-AD	I	**DEMO, TASK, NOTE, SHBM, USAGE**	**DTR**	**0.660**
	J	TASK, SHBM	GBR	0.582
CSE	I	TASK, NOTE, SHBM	GBR	0.738*
	J	**TASK, SHBM**	**GBR**	**0.748*†**
SWLS	I	DEMO, NOTE, SHBM, USAGE	DTR	0.755*
	J	**SHBM, USAGE**	**DTR**	**0.857**†**

* = $p < 0.01$, ** = $p < 0.001$, † = overall assessment experiment configuration across Tables 4-7 with the numerically highest correlation, DTR = decision tree regressor, GBR = gradient-boosted regressor, RFR = random forest regressor.

Using EARLY FUSION, we perform recursive feature elimination with cross validation, principal component analysis, and an autoencoder each as a pre-prediction step. Of the three approaches, only the autoencoder produced improved results, which are presented here. Starting at autoencoder bottleneck size of $m = 25$, we made bottleneck size increments of 50 up to and including $m = 175$. Using this dimensionality reduction technique, the results for four assessments' EARLY FUSION results improved. The improvements and values of *m* for these assessments are shown in Table 6.

Table 6: Autoencoder improved EARLY FUSION results: Numerically highest correlations shown for independent (I) and joint (J) prediction experiment configurations.

Assessment		Marker Sets	Algorithm	*m*	*r*
RBANS	J	TASK	GBR	175	0.671*

WTAR	J	TASK	DTR	25	0.673*†
FAS	I	TASK, NOTE, SHBM	GBR	125	0.616
DFT	J	TASK, NOTE, SHBM	GBR	175	0.871**†
PRMQ	J	NOTE, USAGE	DTR	175	0.862**†
IADL-C	I	TASK, USAGE	DTR	175	0.575
GDS	I	DEMO, NOTE	DTR	175	0.588
QoL-AD	J	SHBM, USAGE	DTR	75	0.762*†
CSE	J	DEMO, SHBM	GBR	75	0.585
SWLS	I	DEMO, TASK, NOTE, SHBM, USAGE	DTR	75	0.576

* = $p < 0.01$, ** = $p < 0.001$, † = overall assessment experiment configuration across Tables 4-7 with the numerically highest correlation, DTR = decision tree regressor, GBR = gradient-boosted regressor, RFR = random forest regressor.

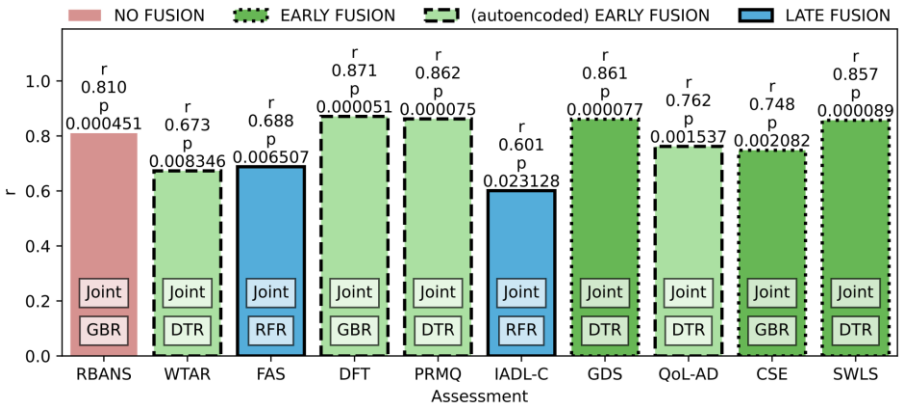
Following EARLY FUSION, we perform leave-one-out cross validation with LATE FUSION. Table 7 reports the numerically highest LATE FUSION correlation results for independent and joint prediction.

Table 7. LATE FUSION results: Numerically highest correlations shown for independent (I) and joint (J) prediction experiment configurations. Bold correlation denotes numerically higher correlation result between I and J rows.

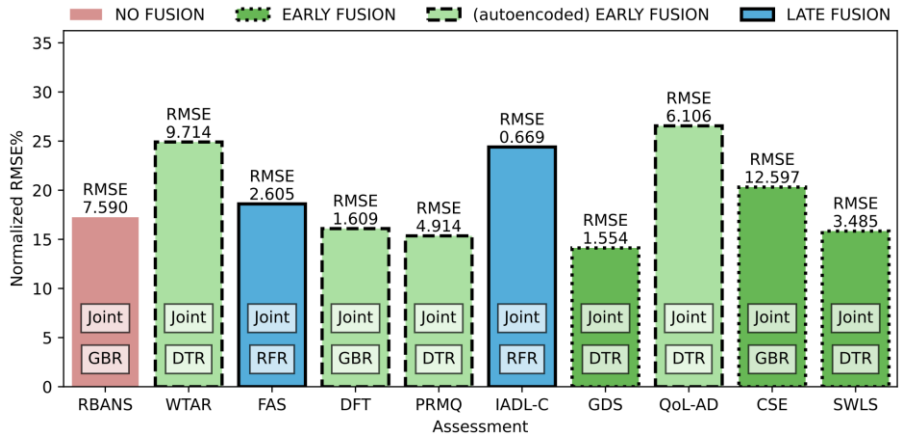
Assessment		**Algorithm**	***r***
RBANS	I	**GBR**	**0.208**
	J	GBR	0.156
WTAR	I	RFR	0.602
	J	**RFR**	**0.607**
FAS	I	RFR	0.580
	J	**RFR**	**0.688*†**
DFT	I	**DTR**	**0.604**
	J	DTR	0.590
PRMQ	I	**DTR**	**0.417**
	J	DTR	0.411
IADL-C	I	GBR	0.484
	J	**RFR**	**0.601†**
GDS	I	**DTR**	**0.812****
	J	DTR	0.775*
QoL-AD	I	**RFR**	**0.249**
	J	RFR	0.153
CSE	I	GBR	0.425
	J	**GBR**	**0.471**
SWLS	I	**DTR**	**0.577**
	J	DTR	0.563

* = $p < 0.01$, ** = $p < 0.001$, † = overall assessment experiment configuration across Tables 4-7 with the numerically highest correlation, DTR = decision tree regressor, GBR = gradient-boosted regressor, RFR = random forest regressor.

To summarize the results across all the experiment configurations reported in Tables 4-7, the correlation coefficient *r* (and associated *p*-value), root mean squared error, and normalized root mean squared error percentage values for each assessment's experiment configuration with the overall numerically highest correlation (experiments denoted with † in Tables 4-7) are shown in Figure 5.



(a) Correlation coefficient (*r*) bars with source *r* and associated *p*-value.



(b) Normalized root mean squared error (RMSE) with source root mean squared error values.

Figure 5. Correlation coefficient (a) and normalized root mean square error (b) bars for each assessment's overall numerically highest correlation experiment configuration in Tables 4-7. DTR = decision tree regressor, GBR = gradient-boosted regressor, RFR = random forest regressor.

5 DISCUSSION

In this paper, we explored multimodal fusion and joint prediction as methods for predicting ten clinical assessments scores (RBANS, WTAR, FAS, DFT, PRMQ, IADL-C, GDS, QoL-AD, CSE, and SWLS) for 14 participants with MCI. From our experimental results in Tables 4-7, we observed widespread prediction

correlations across the assessments, ranging from weak ($r < 0.3$) to strong ($r > 0.5$). Given the relatively small sample size and the experimentation with several variables and configurations, such a range of correlations is expected. The weaker correlations occurred in cases where a clinical assessment exhibited a relatively small variance for the sample, such as WTAR, which had the lowest variance (coefficient of variation = 10.292%) and the second lowest overall prediction correlation of $r = 0.673$ (see Tables 2 and Figure 5). Typically, the lower the variance, the more difficult the distribution is to learn and predict, making it a more challenging clinical health measure to monitor. In the case of WTAR, since it represents an assessment of premorbid abilities, scores are unlikely to be significantly impacted by brain neurodegeneration until late in the process. Weaker correlations can also occur under different configurations (e.g., NO FUSION and independent prediction vs. LATE FUSION with joint prediction), where one configuration yields stronger predictive performance than others, indicating that these computational configurations are important for constructing more predictive, and thus usable, models. For example, IADL-C had the lowest prediction correlation of $r = 0.113$ across all the results tables (Tables 4-7); however, when predicting IADL-C with LATE FUSION and joint prediction, a correlation of $r = 0.601$ was achieved. Overall, we identified configurations for five assessments (RBANS, DFT, PRMQ, GDS, SWLS) that yielded strong prediction correlations ($r > 0.810$; Bonferroni-corrected significant correlations $p < 0.001$). For example, our machine learning configuration for the GDS yielded $r = 0.861$ ($p = 0.000077$) using DEMO, SHBM, and USAGE behavior markers with a decision tree regressor and joint prediction. Given that a depressed state can significantly impact activity level and interests, it may not be surprising that app usage and SHBM behavior markers were able to predict this mood state more easily than a questionnaire that more holistic captures quality of life and includes questions about marriage and financial situation. We also identified the experiment configurations for the other five assessments (WTAR, FAS, IADL-C, QoL-AD, CSE) that produced moderate prediction correlations ($r > 0.601$).

In general, EARLY FUSION was the most accurate experiment setup with four of the numerically highest overall results utilizing EARLY FUSION (RBANS, GDS, CSE, and SWLS), plus an additional four numerically highest results that employed autoencoders (WTAR, DFT, PRMQ, and QoL-AD). RBANS, DFT, PRMQ, GDS, and SWLS are the assessments that were most accurately predicted ($0.810 \leq r \leq 0.871$), while WTAR, FAS, IADL-C, QoL-AD, and CSE were less accurately predicted ($0.601 \leq r \leq 0.762$). These correlations are stronger than previous work using smart home features to predict RBANS ($r = 0.40$), PRMQ ($r = 0.31$), and GDS ($r = 0.21$) [8]. They are also similar to previous work using a fusion of smart home and smartwatch behavior markers with a larger SHiB sample size of N = 21 (RBANS $r = 0.962$, WTAR $r = 0.879$, FAS $r = 0.806$) [5]. Of the different behavior marker sets, TASK appeared to be the marker set with greatest potential for predicting these assessments, with 26 result occurrences in Tables 4-7, followed by SHBM markers (23 result occurrences), followed by USAGE markers (20 result occurrences). NOTE and DEMO sets were the least informative behavior markers, with 13 and 12 result occurrences, respectively. The DEMO set is likely not as informative because the four objective laboratory measures are scaled using age-correction. In support of this statement, 9 of the 12 instances in which DEMO was an informative marker are among the self-report measures. The behavior markers in the NOTE set are probably less prevalent in Tables 4-7 compared to TASK because the NOTE section was taught in the last two weeks of the intervention and was less widely used by the MCI study participants.

Regarding using joint features to improve prediction results, joint prediction did not consistently improve every assessment result, though it did improve results over independent prediction for about half of the

assessments (6/10 NO FUSION, 6/10 EARLY FUSION, 6/10 auto encoded EARLY FUSION, and 4/10 LATE FUSION). Of the overall numerically highest correlation results for each assessment, all ten used joint prediction (see Tables 4-7). These improvements suggest there appears to be predictive utility in using joint features for mapping sensor data to clinical assessment scores, though it is dependent on the target assessment and the predictive assessments that are employed. As for the importance of the algorithm used (decision tree regressor, random forest regressor, or gradient-boosted regressor), decision tree regressor was the most frequent algorithm in Tables 4-7 (35 occurrences), followed by gradient-boosted regressor (26 occurrences), then random forest regressor (9 occurrences). Though these occurrences do not suggest a clear best algorithm for all experiments, decision tree regressor and gradient-boosted regressor seem well suited for EARLY FUSION experiments (5/8 and 3/8 of the overall numerically highest results use decision tree regressor and gradient-boosted regressor in Tables 5 and 6, respectively) and random forest regressor seems well suited for LATE FUSION experiments (both the overall numerically highest FAS and IADL-C experiments use random forest regressor in Table 7). These findings indicate that the choice of algorithm for each configuration is also important and should be selected individually for each assessment. Future work will be needed to identify whether a standard algorithm configuration can be identified that may be best for different types of assessment (e.g., self-report, memory testing, etc.).

5.1 NO FUSION Experiments

NO FUSION trains machine learning algorithms for each behavior marker set individually to determine the most relevant set of behavior markers for each assessment. To account for possible overfitting due to the small SHiB sample, we used leave-one-out cross validation to maximize the amount of training data available. When evaluating DEMO markers' predictive abilities for the self-report measures that were not age-corrected and captured mood, behavior and everyday function, the digital memory notebook and smart home modalities outperformed traditional demographic features. This is because demographic markers do not capture behavior patterns like sensor-based behavior markers can over time. Comparing the digital memory notebook modality to the smart home modality, we observed that three of the ten clinical measures made use of SHBM as the most predictive marker set with NO FUSION (IADL-C, QoL-AD, and SLWS). Of interest, one of the measures best captured by the SHBM (i.e., IADL-C) assesses ability to complete complex activities of daily living (e.g., managing medications, cooking), which the smart home sensors would be in the best position to assess. The remaining self-report and clinical assessments were best predicted using EMMA markers, suggesting that the digital memory notebook markers may provide more predictive power than smart home markers. These observed results may be due to the EMMA app serving as a tool to assist with organization and completion of everyday task and requiring active engagement by the user, as opposed to the smart home modality which is more passive by design.

5.2 EARLY FUSION and LATE FUSION Experiments

In contrast to NO FUSION, we explored two different types of fusion approaches, EARLY FUSION (see Tables 5 and 6 for results) and LATE FUSION (see Table 7 for results), to combine behavior markers from different modalities. For all assessments, fusion improved prediction accuracy; however, RBANS is a unique case because the most predictive EARLY FUSION marker set combination turns out to be TASK markers alone, yielding the same prediction accuracy as gradient-boosted joint prediction with NO FUSION (r = 0.810). This is

a strong correlation (p = 0.000451), implying the tasks and how they are written in EMMA by MCI participants are quite indicative of their current cognitive status (RBANS) as opposed to their premorbid ability (WTAR). In general, EARLY FUSION appeared more appropriate for predicting the assessments utilized in this study because the overall numerically highest correlation result was achieved with EARLY FUSION and EARLY FUSION with autoencoding for eight of the ten assessments (see Tables 5 and 6 and Figure 5). All assessments except RBANS and WTAR benefitted from a fusion of at least two marker sets. Most assessments had the best results utilizing a fusion of two to three marker sets, while QoL-AD and SWLS were the only assessments to utilize all five available marker sets. Assessments trying to capture constructs such as quality of life that measure behavior more holistically, may require more dimensions to accurately assess. Once again, it appeared that EMMA sets are providing the most predictive value, being used in all numerically highest EARLY fusions. SHBM markers appeared in six assessments' numerically highest EARLY FUSION results. Consistent with the NO FUSION results, DEMO and NOTE markers provided the least predictive utility. This suggests that, for the current data with MCI participants, computational complexity can be reduced by eliminating processing of NOTE text.

The application of autoencoders as a feature reduction technique improved results for four of the assessments. This is helpful in overcoming the discrepancy between the number of behavior markers and the number of participants; however, upon investigating the number of reduced features (m) that produces the improved results, m is still quite large. For six assessments, m was in the range of 125 to 175 (see Table 6). Only WTAR had a value close to the number of samples (m = 25). Since autoencoders aim to compress a large feature space, it seems the multimodal feature space is valuable in its many dimensions because larger values of m produced the strongest results for some assessments.

For the LATE FUSION results in Table 7, FAS and IADL-C were the only assessments exhibiting slightly better correlations with LATE FUSION than with EARLY FUSION, with a 11.688% and 4.522% improvement, respectively. On the other hand, some assessments, like RBANS and QoL-AD performed particularly poor under LATE FUSION, perhaps because of the inconsistent accuracy a single marker set-trained algorithm produces. As noted earlier, quality of life assessments (i.e., QoL-AD and SLWS) were the only assessments using all five marker sets in their numerically highest fusion result. Coupling this with QoL-AD's poor LATE FUSION performance, it appears multiple dimensions of behavior and routine from different modalities were necessary to accurately capture similar information contained within a holistic assessment score like QoL-AD.

6 CONCLUSION

In this paper, we explored multimodal fusion and joint prediction of behavior markers for predicting objective (RBANS, WTAR, FAS, DFT) and self-report (PRMQ, IADL-C, GDS, QoL-AD, CSE, and SWLS) clinical scores, for participants with mild cognitive impairment. The modalities we used included smart home environments, a compensatory digital memory notebook iPad app called EMMA, and traditional demographic information. Our smart home system continuously and unobtrusively collected data that was sent to activity recognition algorithms to help model a resident's everyday routine. Furthermore, the smart home system was integrated with the EMMA digital memory notebook app, providing context-aware prompts to help residents navigate their daily routine and remember important information. From these modalities, we extracted behavior markers and explored different fusion techniques to map the markers into clinical assessment scores. Using decision tree, random forest, and gradient boosting regression algorithms, we achieved at least moderate correlations ($r \geq$

0.601) between actual assessment scores and predicted scores. For two of the four objective measures (RBANS and DFT) and three of the six self-report assessments (PRMQ, GDS, and SLWS), we achieved strong correlations ($r \geq 0.810$). Prediction results using joint prediction and multimodal fusion offered improvements over baseline independent predictions for all assessments. Of the behavior marker sets, EMMA task-based markers and smart home-based markers appeared the most informative, suggesting a fusion of multiple modalities, including participant-authored text, may offer the most promising prediction results.

The small sample size (N = 14 participants with both EMMA and smart home data) and the inconsistent use of EMMA for some participants are two limitations of the study. For the former, we anticipate prediction performance would improve with more participants and longer smart home and EMMA data collection periods. For the latter, while participants were trained and encouraged to use the memory app multiple times per day, not all participants used the note and task features enough to provide a large corpus of text for analysis. Future work aims to collect a larger corpus of note and task text from the aging population along a continuum from healthy to mild dementia, explore additional feature selection and reduction techniques for each assessment's overall numerically highest experiment configuration, and utilize EMMA text-based behavior markers for further integration with smartwatch and smart home environments. The latter offers an opportunity for health care providers and caregivers to automatically be notified if a behavior change occurs that may be indicative of health status change.

ACKNOWLEDGMENTS

The authors would like to thank Justin Frow and Sarah Norman and members of the WSU Neuropsychology and Aging Laboratory for their work recruiting study participants, delivering the intervention, and collecting and scoring the clinical data. The authors would also like to thank Jessamyn Dahmen, Eric Chen, Brian Thomas, Aaron Crandall, and members of the WSU CASAS lab for their assistance with designing and implementing the smart home and digital memory notebook app tools.

REFERENCES

[1] S. T. Farias, D. Mungas, B. R. Reed, D. Harvey, D. Cahn-Weiner, and C. Decarli, "MCI is associated with deficits in everyday functioning," Alzheimer Dis. Assoc. Disord., vol. 20, no. 4, pp. 217–223, Dec. 2006, doi: 10.1097/01.wad.0000213849.51495.d9.

[2] N. Raghunath, J. Dahmen, K. Brown, D. Cook, and M. Schmitter-Edgecombe, "Creating a digital memory notebook application for individuals with mild cognitive impairment to support everyday functioning," Disabil. Rehabil. Assist. Technol., vol. 15, no. 4, pp. 421–431, May 2020, doi: 10.1080/17483107.2019.1587017.

[3] J. Dahmen, B. Minor, D. Cook, T. Vo, and M. S. Edgecombe, "Smart home-driven digital memory notebook support of activity self-management for older adults," Gerontechnology, vol. 17, no. 2, pp. 113–125, Aug. 2018, doi: 10.4017/gt.2018.17.2.005.00.

[4] G. Sprint, D. J. Cook, R. Fritz, and M. Schmitter-Edgecombe, "Using smart homes to detect and analyze health events," Computer, vol. 49, no. 11, pp. 29–37, Nov. 2016, doi: 10.1109/MC.2016.338.

[5] D. J. Cook and M. Schmitter-Edgecombe, "Fusing ambient and mobile sensor features into a behaviorome for predicting clinical health scores," IEEE Access, vol. 9, pp. 65033–65043, 2021, doi: 10.1109/ACCESS.2021.3076362.

[6] P. N. Dawadi, D. J. Cook, and M. Schmitter-Edgecombe, "Automated cognitive health assessment from smart home-based behavior data," IEEE J. Biomed. Health Inform., vol. 20, no. 4, pp. 1188–1194, Jul. 2016, doi: 10.1109/JBHI.2015.2445754.

[7] A. Akl, J. Snoek, and A. Mihailidis, "Unobtrusive detection of mild cognitive impairment in older adults through home monitoring," IEEE J. Biomed. Health Inform., vol. 21, no. 2, pp. 339–348, Mar. 2017, doi: 10.1109/JBHI.2015.2512273.

[8] A. Alberdi et al., "Smart home-based prediction of multidomain symptoms related to alzheimer's disease," IEEE J. Biomed. Health Inform., vol. 22,

no. 6, pp. 1720–1731, Nov. 2018, doi: 10.1109/JBHI.2018.2798062.

[9] K. C. Fraser, K. Lundholm Fors, M. Eckerström, F. Öhman, and D. Kokkinakis, “Predicting MCI status from multimodal language data using cascaded classifiers,” Front. Aging Neurosci., vol. 11, 2019, doi: 10.3389/fnagi.2019.00205.

[10] M. Schmitter-Edgecombe et al., “Partnering a compensatory application with activity-aware prompting to improve use in individuals with amnestic mild cognitive impairment: a randomized controlled pilot clinical trial,” J. Alzheimers Dis., vol. 85, no. 1, pp. 73–90, Jan. 2022, doi: 10.3233/JAD-215022.

[11] S. Aminikhanghahi and D. J. Cook, “Enhancing activity recognition using CPD-based activity segmentation,” Pervasive Mob. Comput., vol. 53, pp. 75–89, Feb. 2019, doi: 10.1016/j.pmcj.2019.01.004.

[12] B. Minor, J. R. Doppa, and D. J. Cook, “Data-driven activity prediction: algorithms, evaluation methodology, and applications,” in Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2015, pp. 805–814. doi: 10.1145/2783258.2783408.

[13] P. N. Dawadi, D. J. Cook, and M. Schmitter-Edgecombe, “Modeling patterns of activities using activity curves,” Pervasive Mob. Comput., vol. 28, pp. 51–68, Jun. 2016, doi: 10.1016/j.pmcj.2015.09.007.

[14] G. Sprint, D. J. Cook, and R. Fritz, “Behavioral differences between subject groups identified using smart homes and change point detection,” IEEE J. Biomed. Health Inform., vol. 25, no. 2, pp. 559–567, Feb. 2021, doi: 10.1109/JBHI.2020.2999607.

[15] A. R. Javed et al., “Automated cognitive health assessment in smart homes using machine learning,” Sustain. Cities Soc., vol. 65, p. 102572, Feb. 2021, doi: 10.1016/j.scs.2020.102572.

[16] 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. Transl. Eng. Health Med., vol. 4, pp. 1–11, 2016, doi: 10.1109/JTEHM.2016.2579638.

[17] F. S. Aronsson, M. Kuhlmann, V. Jelic, and P. Östberg, “Is cognitive impairment associated with reduced syntactic complexity in writing? Evidence from automated text analysis,” Aphasiology, vol. 0, no. 0, pp. 1–14, Apr. 2020, doi: 10.1080/02687038.2020.1742282.

[18] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection of depression-related posts in reddit social media forum,” IEEE Access, vol. 7, pp. 44883–44893, 2019, doi: 10.1109/ACCESS.2019.2909180.

[19] A. H. Alkenani, Y. Li, Y. Xu, and Q. Zhang, “Predicting Alzheimer’s disease from spoken and written language using fusion-based stacked generalization,” J. Biomed. Inform., vol. 118, p. 103803, Jun. 2021, doi: 10.1016/j.jbi.2021.103803.

[20] L. K. Dickerson et al., “Language impairment in adults with end-stage liver disease: application of natural language processing towards patient-generated health records,” Npj Digit. Med., vol. 2, no. 1, Art. no. 1, Nov. 2019, doi: 10.1038/s41746-019-0179-9.

[21] J. Bullard, C. Ovesdotter Alm, X. Liu, Q. Yu, and R. Proaño, “Towards early dementia detection: fusing linguistic and non-linguistic clinical data,” in Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology, San Diego, CA, USA, Jun. 2016, pp. 12–22. doi: 10.18653/v1/W16-0302.

[22] Q. He, B. P. Veldkamp, C. A. W. Glas, and S. M. van den Berg, “Combining text mining of long constructed responses and item-based measures: a hybrid test design to screen for posttraumatic stress disorder (PTSD),” Front. Psychol., vol. 10, 2019, doi: 10.3389/fpsyg.2019.02358.

[23] F. Ö. Sönmez and Y. Maleh, “Prediction of satisfaction with life scale using linguistic features from Facebook status updates: smart life,” in Machine Intelligence and Data Analytics for Sustainable Future Smart Cities, U. Ghosh, Y. Maleh, M. Alazab, and A.-S. K. Pathan, Eds. Cham: Springer International Publishing, 2021, pp. 119–144. doi: 10.1007/978-3-030-72065-0_8.

[24] L. Calzà, G. Gagliardi, R. Rossini Favretti, and F. Tamburini, “Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia,” Comput. Speech Lang., vol. 65, p. 101113, Jan. 2021, doi: 10.1016/j.csl.2020.101113.

[25] F. Bertini, D. Allevi, G. Lutero, D. Montesi, and L. Calzà, “Automatic speech classifier for mild cognitive impairment and early dementia,” ACM Trans. Comput. Healthc., vol. 3, no. 1, p. 8:1-8:11, Oct. 2021, doi: 10.1145/3469089.

[26] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, “Multimodal fusion of BERT-CNN and gated CNN representations for depression detection,” in Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, New York, NY, USA, Oct. 2019, pp. 55–63. doi: 10.1145/3347320.3357694.

[27] R. Ostrand and J. Gunstad, “Using automatic assessment of speech production to predict current and future cognitive function in older adults,” J.

Geriatr. Psychiatry Neurol., p. 0891988720933358, Jul. 2020, doi: 10.1177/0891988720933358.

[28] M. Martinc and S. Pollak, “Tackling the adress challenge: a multimodal approach to the automated recognition of Alzheimer’s dementia,” in Interspeech 2020, Oct. 2020, pp. 2157–2161. doi: 10.21437/Interspeech.2020-2202.

[29] M. Gónzalez Atienza, J. A. González López, and A. M. Peinado, “An automatic system for dementia detection using acoustic and linguistic features,” Jan. 2021, Accessed: Jun. 07, 2021. [Online]. Available: https://digibug.ugr.es/handle/10481/66645

[30] R. Haulcy and J. Glass, “Classifying Alzheimer’s disease using audio and text-based representations of speech,” Front. Psychol., vol. 11, p. 624137, Jan. 2021, doi: 10.3389/fpsyg.2020.624137.

[31] F. T. Giuntini, M. T. Cazzolato, M. de J. D. dos Reis, A. T. Campbell, A. J. M. Traina, and J. Ueyama, “A review on recognizing depression in social networks: challenges and opportunities,” J. Ambient Intell. Humaniz. Comput., vol. 11, no. 11, pp. 4713–4729, Nov. 2020, doi: 10.1007/s12652-020-01726-4.

[32] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, “Alzheimer’s dementia recognition through spontaneous speech: the ADReSS challenge,” ArXiv200406833 Cs Eess Stat, Aug. 2020, Accessed: Feb. 13, 2022. [Online]. Available: http://arxiv.org/abs/2004.06833

[33] C. Dreisbach, T. A. Koleck, P. E. Bourne, and S. Bakken, “A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data,” Int. J. Med. Inf., vol. 125, pp. 37–46, May 2019, doi: 10.1016/j.ijmedinf.2019.02.008.

[34] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, “Identifying mild cognitive impairment and mild Alzheimer’s disease based on spontaneous speech using asr and linguistic features,” Comput. Speech Lang., vol. 53, pp. 181–197, Jan. 2019, doi: 10.1016/j.csl.2018.07.007.

[35] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: a survey and taxonomy,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

[36] G. L. Sprint, D. J. Cook, and D. L. Weeks, “Patient similarity and joint features for rehabilitation outcome prediction,” Proc. 2016 Workshop Knowl. Discov. Healthc. KDH IJCAI, p. 7, 2016.

[37] L. A. Chudoba, A. S. Church, J. B. Dahmen, K. D. Brown, and M. Schmitter-Edgecombe, “The development of a manual-based digital memory notebook intervention with case study illustrations,” Neuropsychol. Rehabil., vol. 30, no. 9, pp. 1829–1851, Oct. 2020, doi: 10.1080/09602011.2019.1611606.

[38] D. J. Cook, A. S. Crandall, B. L. Thomas, and N. C. Krishnan, “CASAS: a smart home in a box,” Computer, vol. 46, no. 7, pp. 62–69, Jul. 2013, doi: 10.1109/MC.2012.328.

[39] R. C. Petersen et al., “Criteria for mild cognitive impairment due to Alzheimer’s disease in the community,” Ann. Neurol., vol. 74, no. 2, pp. 199–208, Aug. 2013, doi: 10.1002/ana.23931.

[40] 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., vol. 20, no. 3, pp. 310–319, Jun. 1998, doi: 10.1076/jcen.20.3.310.823.

[41] The Psychological Corporation, “Wechsler test of adult reading,” San Antonio, TX, 2001.

[42] D. Delis, E. Kaplan, and J. Kramer, Delis-Kaplan Executive Function System: Examiner’s Manual. The Psychological Corporation. Accessed: Sep. 10, 2021. [Online]. Available: https://www.pearsonclinical.co.uk/Psychology/AdultCognitionNeuropsychologyandLanguage/AdultAttentionExecutiveFunction/Delis-KaplanExecutiveFunctionSystem(D-KEFS)/Delis-KaplanExecutiveFunctionSystem(D-KEFS).aspx

[43] G. Smith, S. D. Sala, R. H. Logie, and E. A. Maylor, “Prospective and retrospective memory in normal ageing and dementia: A questionnaire study,” Memory, vol. 8, no. 5, pp. 311–321, Sep. 2000, doi: 10.1080/09658210050117735.

[44] M. Schmitter-Edgecombe, C. Parsey, and R. Lamb, “Development and psychometric properties of the instrumental activities of daily living: compensation scale,” Arch. Clin. Neuropsychol., vol. 29, no. 8, pp. 776–792, Dec. 2014, doi: 10.1093/arclin/acu053.

[45] J. I. Sheikh and J. A. Yesavage, “Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version,” Clin. Gerontol. J. Aging Ment. Health, vol. 5, no. 1–2, pp. 165–173, 1986, doi: 10.1300/J018v05n01_09.

[46] R. G. Logsdon, L. E. Gibbons, S. M. McCurry, and L. Teri, “Quality of life in Alzheimer’s disease: patient and caregiver reports,” J. Ment. Health Aging, vol. 5, no. 1, pp. 21–32, 1999.

[47] M. A. Chesney, T. B. Neilands, D. B. Chambers, J. M. Taylor, and S. Folkman, “A validity and reliability study of the coping self-efficacy scale,” Br. J.

Health Psychol., vol. 11, no. 3, pp. 421–437, 2006, doi: 10.1348/135910705X53155.

[48] E. Diener, R. A. Emmons, R. J. Larsen, and S. Griffin, “The satisfaction with life scale,” J. Pers. Assess., vol. 49, no. 1, pp. 71–75, Feb. 1985, doi: 10.1207/s15327752jpa4901_13.

[49] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, “The development and psychometric properties of LIWC2015,” 2015, doi: 10.15781/T29G6Z.

[50] B. Shickel, S. Siegel, M. Heesacker, S. Benton, and P. Rashidi, “Automatic detection and classification of cognitive distortions in mental health text,” in 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), Oct. 2020, pp. 275–280. doi: 10.1109/BIBE50027.2020.00052.

[51] R. Wang et al., “Tracking depression dynamics in college students using mobile phone and wearable sensing,” Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 2, no. 1, p. 43:1-43:26, Mar. 2018, doi: 10.1145/3191775.

[52] L. Wu, S. L. Oviatt, and P. R. Cohen, “Multimodal integration-a statistical view,” IEEE Trans. Multimed., vol. 1, no. 4, pp. 334–341, Dec. 1999, doi: 10.1109/6046.807953.

[53] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” Artif. Intell., vol. 97, no. 1–2, pp. 273–324, 1997, doi: 10.1016/S0004-3702(97)00043-X.

[54] Y. Wang, H. Yao, and S. Zhao, “Auto-encoder based dimensionality reduction,” Neurocomputing, vol. 184, pp. 232–242, Apr. 2016, doi: 10.1016/j.neucom.2015.08.104.

[55] R. Blagus and L. Lusa, “Gradient boosting for high-dimensional prediction of rare events,” Comput. Stat. Data Anal., vol. 113, pp. 19–37, Sep. 2017, doi: 10.1016/j.csda.2016.07.016.