

Applications of Machine Learning to Support Dementia Care through Commercially Available Off-the-Shelf Sensing

George Netscher



Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2016-204

<http://www2.eecs.berkeley.edu/Pubs/TechRpts/2016/EECS-2016-204.html>

December 15, 2016

Copyright © 2016, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

**Applications of Machine Learning to Support Dementia Care through
Commercially Available Off-the-Shelf Sensing**

by

George Netscher

A project report submitted in partial satisfaction of the
requirements for the degree of
Master of Science, Plan II

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Alexandre M. Bayen, Research Advisor
Trevor Darrell, Second Reader

Fall 2016

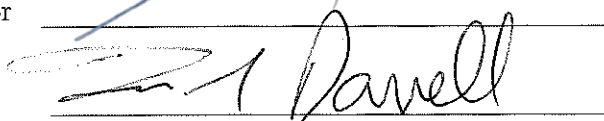
The project report of George Netscher, titled Applications of Machine Learning to Support Dementia Care through Commercially Available Off-the-Shelf Sensing, is approved:

Research Advisor



Date 12/13/16

Second Reader



Davell

Date 12/12/16

University of California, Berkeley

**Applications of Machine Learning to Support Dementia Care through
Commercially Available Off-the-Shelf Sensing**

Copyright 2016
by
George Netscher

Abstract

Applications of Machine Learning to Support Dementia Care through Commercially Available Off-the-Shelf Sensing

by

George Netscher

Master of Science, Plan II in Computer Science

University of California, Berkeley

Alexandre M. Bayen, Research Advisor

In this report, we discuss a project beginning August 2014 and ending in December 2016 through which four applications of machine learning to dementia care were explored. The purpose of this project was to determine how advances in machine learning could be applied to commercially available off-the-shelf sensing equipment to make a positive impact in care for individuals with Alzheimer's disease and related dementias (ARD), a cause personally important to the author and the research advisor. The project will be discussed for an audience familiar with the state-of-the-art in machine learning but unfamiliar with the open problems in dementia care. The first chapter gives background on Alzheimer's disease and the context for the current work in terms of the current challenges faced by the Alzheimer's research community. The following four chapters each discuss one application. The first discusses how a wearable system can be designed to support daily monitoring of individuals affected by Alzheimer's disease to study functional changes which can occur as the disease progresses. The second discusses how analysis of speech can be used to detect the presence of dementia. The third discusses how video monitoring can be used to detect safety-critical events with a particular focus on falls. The fourth provides preliminary pilot study results from the application of video monitoring in one 40-resident memory care community. The final chapter concludes by discussing the gaps between the available technology and the current needs and poses suggestions for future work to bridge the gaps.

Contents

Contents	i
List of Figures	iii
List of Tables	vi
1 Introduction and Background	1
1.1 Background on Alzheimer’s Disease	1
1.2 Context of Work	2
2 Functional Monitoring through Wearables	7
2.1 Chapter Abstract	7
2.2 Introduction	8
2.3 The Dementia Care Ecosystem	9
2.4 System Architecture	10
2.5 Biggest Development Challenges	14
2.6 Analysis Methods	16
2.7 Indoor Positioning	19
2.8 Results of Beta Test	23
2.9 Conclusion	26
3 Diagnosis through Speech	29
3.1 Chapter Abstract	29
3.2 Introduction	29
3.3 Background and Related Work	30
3.4 Feature Extraction Process	32
3.5 Classification Process	37
3.6 Results	39
3.7 Discussion	43
3.8 Conclusion	46
4 Fall Detection through Video Analysis	47
4.1 Chapter Abstract	47

4.2	Introduction	47
4.3	Methods	48
4.4	Results and Discussion	52
4.5	Conclusion	56
5	Fall Reduction through Video Review	59
5.1	Chapter Abstract	59
5.2	Introduction	59
5.3	Materials and Methods	61
5.4	Results	63
5.5	Discussion	64
5.6	Conclusions	64
6	Conclusions	65
6.1	Review of Project Report	65
6.2	Final Conclusion: Hybrid Solutions are Required for Practical Challenges . .	66
	Bibliography	68

List of Figures

1.1	Alzheimers disease is the most expensive disease in the US and the only disease in the top six for which the number of deaths is increasing [2]	1
1.2	Factors which impact the likelihood of cognitive decline and dementia	4
2.1	The architecture deployed for the Dementia Care Ecosystem.	10
2.2	The five system components and their uses.	12
2.3	CPU usage before optimizing for battery life. Waking up the CPU frequently to sample from the sensors caused rapid battery loss.	14
2.4	Outliers are those points labeled by DBSCAN as not belonging to any cluster [53].	17
2.5	Global outlier detection with k -NN fails for a nonstationary distribution. Local outlier detection methods provide empirically worse performance as described in described in [29]. We instead prefer the $k - NN$ with a rolling window when anomaly detection over a nonstationary window is required.	19
2.6	RANSAC fits a regression line by consensus, providing robustness to noise [53]. .	20
2.7	KLMS outperforms linear LMS in fitting nonlinear functions when the nonlinear function class is known a priori [45].	22
2.8	The random forest increases classifier accuracy by averaging over random decision trees to reduce variance [53].	24
2.9	Salesforce user interface for care team navigators to view metrics and analysis for monitored individuals with dementia	25
2.10	The Android user interface for administrators and a selected screen from home setup. The interface for non-administrator users provides a subset of the functions available.	26
2.11	Representative plot of room inference from raw RSSI	27
2.12	Representative plot of inferred room location with outlier detection applied . . .	27
2.13	Representative plot of step count data with trend detection applied	28

3.1	Distribution of the 126 individuals with respect to disease. The set comprises 66 <i>healthy controls</i> (HC, 52.4%), and 60 individuals with <i>Alzheimer's disease and related Dementias</i> (ADRD, 47.6%). Of the affected individuals, the primary diagnosis for 16 is <i>Alzheimer's disease</i> (AD, 12.7%), for 20 is <i>behavioral Frontotemporal Dementia</i> (bvFTD, 15.9%), for 1 is <i>Dementia with Lewy Bodies</i> (DLB, 0.8%), for 23 is <i>Primary Progressive Aphasia</i> (PPA, 18.3%). Within the PPA segment, 14 show the <i>semantic variant</i> (svPPA, 11.1%), 7 show the <i>right semantic variant</i> (rsvPPA, 5.6%), and 2 show the <i>non-fluent variant</i> (nfvPPA, 1.6%) . . .	34
3.2	Distribution of the 126 individuals with respect to gender and age.	34
3.3	Vocabulary Features Extraction Process. Each word considered relevant is used to compute aggregation functions. The others are counted to compute the relevance ratio	36
3.4	Procedure to perform classification using regressors. 0, 1 and 2 represent the labels to predict (HC, AD, FTD, PPA). One regressor is associated to each of them	38
3.5	The best results in determining whether a speech segment belongs to an individual with dementia or a healthy control. Two-step AdaBoost, or Selective Boosting, demonstrates 92% accuracy and greater than 90% precision and recall.	39
3.6	The best results in determining whether a speech segment belongs to an individual with dementia or a healthy control. Most tree-based methods reach accuracies higher than 80%.	40
3.7	The best results in determining the diagnosis if present. Gradient Boosting demonstrates 70% accuracy, greatly benefitted from high recall among healthy controls.	40
3.8	The effect of feature selection on the bimodal classification results. Feature selection is performed a priori using the AdaBoost score function, so the change in accuracy for AdaBoost is most indicative.	41
3.9	The best results when limited to 15 features. Note multiBoost still demonstrates 85% accuracy. Here feature selection is performed using a decision tree with Gini criterion.	41
4.1	Examples of data from the day-time and night-time settings.	49
4.2	Domain confusion net, based on [74], used for experiments. Note that the first seven layers are initialized from the VGG weights [68]. We lock the weights for all layers except fc7 and fc8. In implementation, we use two fcD layers with shared weights to connect to light and dark fc7 layers, respectively.	50
4.3	An example of deep artistic style transfer from [26] whereby the content of image A is transformed into the style of 3 separate paintings in images B, C, and D. . .	53
4.4	Examples of style transfer with originals on the left and transformed images on the right.	54

4.5	Measurement of the domain classifier’s ability to distinguish between light and dark domains over training process. There are initial spikes in precision and recall, followed by convergence to under 50%. Note that the light and dark domain confusion net results in Tables 4.2 and 4.3 occur at 25,000 and 15,000 iterations, respectively.	55
4.6	Success and failure examples in fall detection in the light domain.	57
4.7	Success and failure examples in fall detection in the dark domain.	58
5.1	Equipment. IP cameras were placed in all common areas and approved private rooms. Video was transmitted from the cameras to the network attached storage (NAS) via Wi-Fi where it was maintained locally for 72 hours after which it was transmitted to a remote server for archiving. Live video and video from the previous 72 hours were made available to facility management via smartphone applications.	61
5.2	Fall rate. In the four months prior to video review, the fall rate at the community was 10.5 2.5 falls per month, 79% of the national average. In the final month, 2 falls occurred, 17% of the national average.	63

List of Tables

3.1	Top 10 features for bimodal classification. The symbol \circ denotes composition of functions.	42
3.2	Top 10 features for bimodal classification with ANOVA. The symbol \circ denotes composition of functions.	43
3.3	Top 10 features for multimodal classification. The symbol \circ denotes composition of functions.	43
4.1	SGD solver parameters used to train all nets.	50
4.2	Fall detection results for baseline, domain confusion, and style transfer methods.	51
4.3	Dark domain detection results for domain confusion method. The dark domain detection results in this table correspond to the snapshots used to evaluate fall detection in Table 4.2. Note that the test set used for dark domain detection is the same test set used in Table 4.2 but is partitioned by domain rather than by category.	51

Acknowledgments

This work would not have been possible without help from so many people. I want to thank my advisor Alex Bayen for supporting each of the directions followed as we searched for the best way to make a positive impact and gracefully supporting transitions between many projects and roles. I want to thank my coworker Julien Jacquemot for all of the late nights and long hours working together on projects. I want to thank the many students and faculty who worked on each of these projects and without whom this work would not have been possible including: Pulkit Agrawal, Nick Boyd, Sbastien Levy, Bradley Zylstra, Yanrong Li, Ludovic Thea, Jun Jie Ng, Marie Douriez, Chong Wee Tan, Cyril Tamraz, Peter Pressman, Bob Levenson, Katrin Schenk, Steve Bonasera, Kate Possin, and Bruce Miller. I want to thank each of my office mates for supporting this work through daily discussions and weekend barbecues including Cathy Wu, Jerome Thai, Walid Krichene, and Francois Belletti. Finally, I am so thankful to my girlfriend, Casey Maas. Without her daily support, I am quite sure my sanity would have been lost somewhere along the way.

Chapter 1

Introduction and Background

1.1 Background on Alzheimer's Disease

In the US, Alzheimer's disease is the sixth most common disease and the single most expensive disease (\$236B direct costs; estimated \$221B indirect). As shown in figure 1, among the top six diseases, it is the only disease for which the number of deaths is increasing. As the median age of the US population continues to increase, Alzheimer's disease will only becoming more prevalent, and the resulting cost of Alzheimer's care will continue rising to unsustainable levels [2]. Unfortunately, the drug failure rate for Alzheimer's disease still remains among the highest – currently 99.6% (as compared to 81% for cancer) [14] due to our limited understanding of the brain and the root causes of Alzheimer's disease.

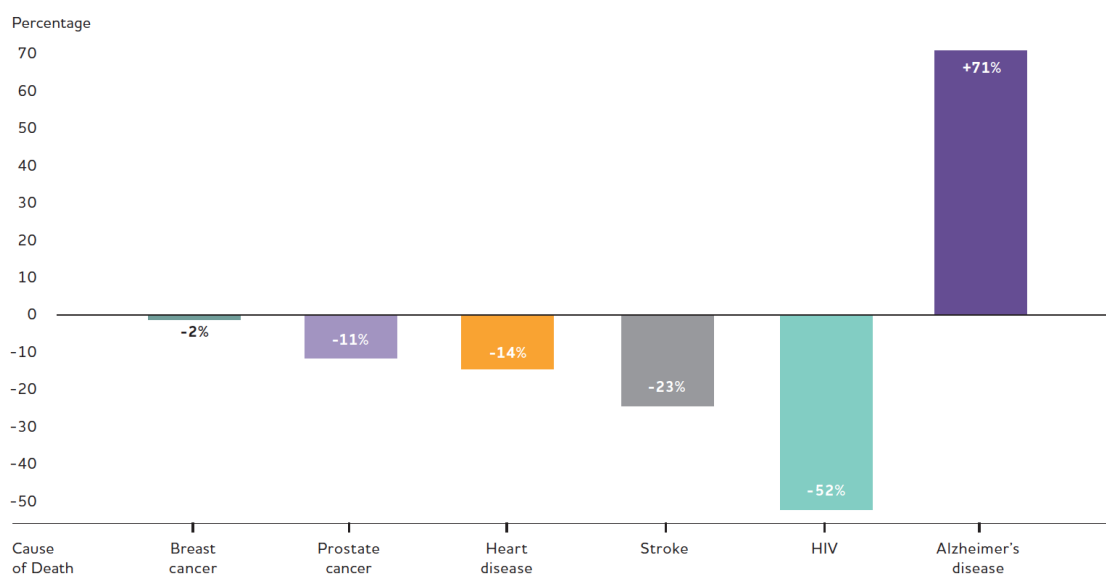


Figure 1.1: Alzheimers disease is the most expensive disease in the US and the only disease in the top six for which the number of deaths is increasing [2]

Alzheimer's disease is characterized by a progressive loss of cognitive ability. In the mild stage, affected individuals may have trouble remembering words, planning, and completing tasks independently. In the moderate stage, those affected might forget their own personal history, become confused about time and place, and suffer alterations in personal behavior and self-control. By the severe stage, individuals usually require full-time care to help complete the activities of daily living (ADLs) such as eating and toileting, and they may experience changes in physical ability such as the ability to walk, sit, and swallow.

The incidence rate of Alzheimer's disease is alarming. 5.4 million people in the United States are affected including 1 in 9 people over 65 years old and 1 in 3 over 85 years old. Thus, almost surely, every person who reads this statement will be personally affected through a loved one, a close personal friend, or a personal diagnosis. One reason for the high incidence rate is that it is not uncommon for an individual with Alzheimers disease to live for 20 years after diagnosis. In much of this time, individuals will require support with activities of daily living and struggle with preventable emergency room visits due to urinary tract infections, extrinsic fall incidents, and bedsores. These incidents could be prevented respectively through proper hydration and toileting habits, proper design of environment to remove external factors contributing to falls, and through periodic changes in body positioning. With costs already at unsustainable levels, Alzheimers disease threatens to cripple the US healthcare system, and Alzheimers disease only represents two thirds of all those affected by dementia [2].

1.2 Context of Work

In this work, we discuss three approaches currently under development in the research community to address the challenges presented by Alzheimer's disease and related dementias.

1. **Curing, delaying, or mitigating disease effects:** This research area focuses broadly on the root cause of the disease. It includes the many pharmaceutical approaches attempted to cure Alzheimer's and many public health studies aimed at determining if certain interventions such as proper diet and exercise can mitigate the effects on a population level.
2. **Early detection:** This research focuses on identifying relevant biomarkers which can be detected before significant brain damage has occurred. By detecting these warning signs early, the available interventions identified in the previous research area can be applied to delay and/or mitigate the effects of the disease.
3. **Caring for those currently affected:** This research area focuses on improving the quality of life and reducing the cost of care for those currently affected. Major themes in this area include reducing the rate of hospitalization where falls are the greatest contributor and enabling individuals to remain independent for longer.

Although several other interesting areas of Alzheimer’s research exist, these three themes capture major thrusts within the research community. Moreover, we focus here because in each area, there appear to be significant opportunities for the development of technology which could make a far-reaching positive impact.

Curing, Delaying, or Mitigating the Effects

The first approach where technology may provide support is in curing, delaying, or mitigating the effects of Alzheimer’s disease. Unfortunately, it seems every year a promising pharmaceutical appears to provide hope before failing in clinical trial. Most recently, Eli Lilly’s experimental Alzheimer’s therapy solanezumab which showed potential for slowing the effects of cognitive impairment in 2015 [55] failed large scale clinical trial in November 2016 [10]. This therapy targets the amyloid plaques that appear as tangles in the brain of those affected by Alzheimer’s disease. The failure of this therapy adds supporting evidence that these amyloid plaques which are characteristic of Alzheimer’s disease may only be a symptom and not the root cause of the disease.

Although research with respect to finding a cure has so far proven unsuccessful, interesting results have been uncovered with respect to delaying the symptoms. For instance, with respect to brain training games, a consensus statement was released by a group of leading geriatricians expressing concerns regarding the lack of supporting evidence [3]. Specifically, although evidence existed that individuals continued to perform well on brain training games, this success did not appear to extend to more general cognitive abilities. Since then, several interesting results have been released including [60] and [4], most notably showing from a sample of 2,832 volunteers that those who completed gamified training sessions were 48% less likely to be diagnosed with some form of dementia after ten years.

The Alzheimer’s Association official stance on preventative measures is based on a 2015 review of the literature [9] in which sufficient evidence was found to support that regular physical exercise and management of cardiovascular risk factors including diabetes, obesity, smoking, and hypertension reduce the risk of cognitive decline and may reduce the risk of dementia. Healthy diet and lifelong learning or cognitive training may also reduce the risk of cognitive decline but sufficient evidence does not exist to suggest that they may reduce the risk of dementia. Our work with respect to this first thrust of Alzheimer’s research is discussed in Chapter 2 where we discuss the design and implementation of a system for monitoring the disease progression of individuals with Alzheimer’s. The aim of this system is to provide researchers with the necessary tools to define more fine-grained relationships between the amount of exercise individuals undertake and the eventual onset of the disease. Traditional approaches require participants to fill out daily surveys and off-the-shelf tools provide limited functionality such as a step count. We design here a open-source platform based on off-the-shelf components for interacting with sensors around the home (e.g., to see if the stove is on) and perform typical signal processing and machine learning techniques on the fine-grained data (e.g., to detect anomalies). More details on similar approaches

are discussed in Chapter 2 alongside the relevant sensing equipment and available machine learning methods.

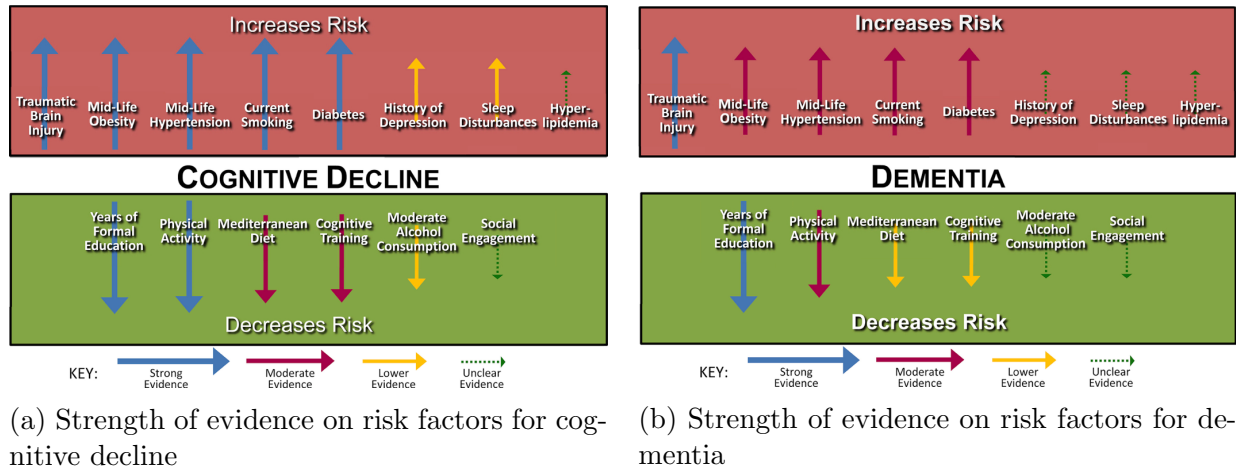


Figure 1.2: Factors which impact the likelihood of cognitive decline and dementia

Early Detection

The second research thrust focuses on identifying ways in which Alzheimer’s disease can be detected before noticeable changes in behavior occur. Typically, Alzheimer’s disease is diagnosed after a family member or friend notices perceptible changes in the individual’s memory. Unfortunately, these changes often become apparent only after significant brain damage has occurred. At this point, the damage is thought to be irreversible [1]. Thus, significant interest exists in the development of screening tools which could be applied early and with high sensitivity to detect individuals living in the community which may have Alzheimer’s disease or related dementia before significant damage occurs. The screening tool would then refer the individual to clinical personnel for more accurate diagnosis.

Clinical diagnosis of Alzheimer’s disease is typically performed through personal history, family history, memory tests including the commonly used mini-mental state exam [22], physical tests such as blood and urine analysis, and brain scans such as CT, MRI, and PET. Although brain scans would seem to be a useful tool for a human expert to perform diagnosis of brain abnormalities, the traditional role of these scans has been to rule out other possible causes of symptoms such as tumor growth [1], [38]. Even with all of these tools, the diagnostic accuracy by clinical experts is surprisingly low. A 5-year review of 919 subjects from Alzheimer’s Disease Centers sponsored by the National Institute on Aging who had died and been autopsied revealed sensitivity for the centers ranging from 70.9% to 87.3% and specificity ranging from 44.3% to 70.8%. Currently, the only way to identify Alzheimer’s disease with 100% accuracy is through post-mortem histology to identify the plaques characteristic of the disease. These low accuracies in diagnoses are particularly startling after

considering the years of expertise required by the clinical expert and the potential harm that can be caused by misdiagnosis. Individuals with Dementia with Lewy Bodies, for instance, respond particularly negatively and can die when inappropriately prescribed anti-psychotic medication [47].

These low diagnostic accuracies suggest that clinicians do not have the necessary tools to obtain clear signals regarding patient state. Thus, a primary focus in this research area is identifying relevant biomarkers that provide strong indications of particular diseases and the stage of these disease. The most relevant work in this regard is with respect to new techniques for identifying Alzheimer’s disease from cerebral spinal fluid [27], [67] where the concentration of amyloid- β -derived diffusible ligands provides a relevant marker for Alzheimer’s disease. In Chapter 3, we discuss how traditional machine learning techniques can be applied to conversational speech data from individuals with Alzheimer’s disease and their caregivers to detect the presence of dementia. This approach is tempting in that a smartphone application could easily be developed to act as an early screening tool, but refining the approach presents difficulties in longitudinal data collection which were not practically feasible within the scope of this project report.

Caring for Those Currently Affected

The final research area focuses on how we can support care for those currently affected by Alzheimer’s disease and related dementias. Particular areas of interest include how we can improve the quality of care and reduce the cost of providing care. The biggest contributors to cost include the need for assistance with the activities of daily living and the high rate of hospitalization for individuals affected by Alzheimer’s disease [2]. A particular focus area in this regard is delaying or reducing the need for institutionalized care by empowering family caregivers and home care services to support care in the home of the affected individual for longer. Work in this area includes the development of proper tools such as those provided by the Alzheimer’s Association and Family Caregiver Alliance for educating caregivers about the resources available to them such as adult day care services and memory care communities which may provide short-term respite care. Another interesting avenue includes the study of how technology can be used with a human assistant in the loop. This includes work done by the UC San Francisco and University of Nebraska Medical Centers on the Dementia Care Ecosystem where anomalies can be detected by home sensors and screened by a low-cost case manager before escalating to the need for an emergency room visit as discussed in Chapter 2.

Several interesting commercial product offerings also exist in this space including traditional fall detection pendants like the Phillips Lifeline, Emerald non-wearable fall detection, and wander detection systems like the GPS SmartSole and Bluetooth SafeWander. Unfortunately, although these products provide wander and fall detection, there are no systems with significant supporting evidence for reducing the rate at which these safety accidents occur. In Chapters 4 and 5, we discuss methods for detecting falls from video and the first deployment of such techniques. Falls are the leading cause of hospitalization in Alzheimer’s care and are

a particular concern in managed care where residents with dementia have been observed to fall at an average rate of 4 times per year, roughly twice that of cognitively healthy elderly residents [17]. Moreover, less than 10% of falls lead to serious injury [15], [17], but 50-75% of elderly fallers experience repeat falls. Although preliminary, it appears the use of cameras in dementia care communities may provide significant benefit with respect to lowering the rate of repeat falls. In one pilot study with a 40-resident memory care community, the fall rate was reduced by 80% following video review of fall incidents after which, personalized changes could be made to individual room environments based on the way in which residents were falling. The technology behind this video fall detection is discussed in Chapter 4 and the results from the pilot study are discussed in 5.

Chapter 2

Functional Monitoring through Wearables

2.1 Chapter Abstract

The increasing availability of wearable computing opens up new avenues for cyberphysical systems which can provide content personalized to user preferences. In the past, testing these ideas has required expertise in disparate areas ranging from embedded systems to machine learning. Max is a platform for rapidly prototyping new ideas in personalized wearable computing without the need for in-depth expertise and long design cycles. It is built from off-the-shelf components – Bluetooth home sensors, an Android smartwatch, and Android smartphone – and (mostly) open source libraries – Android OS and Sci-Kit Learn. From these components, Max is a full-stack system including methods for collecting data from the individual via the watch and from the environment via the sensors; maintaining data securely; transmitting data to a backend server; performing standard machine learning and signal processing tasks such as filtering, classifying, detecting trends, and flagging anomalies; and displaying data through Android UX and Salesforce API. In addition, novel methods for cost-effective approximate indoor positioning are developed. We show one use case for Max, monitoring individuals with Alzheimer’s disease (AD) through the Dementia Care Ecosystem. The Dementia Care Ecosystem defines a new proactive model from the UCSF and UNMC medical centers aimed at reducing emergency room use. This article describes the design and implementation of Max including the challenges faced, the tradeoffs made, and beta test results from 13 healthy users over 39 total months. These results show 96.1% accuracy in room-locationing and many trade-offs that must be made concerning battery life.

2.2 Introduction

In the last four years the global market for wearable devices has grown from \$0.75B to \$2.93B. This nearly 400% growth has been fueled mainly by high market demand for fitness trackers which can monitor bodily signals such as step count, heart rate, and hours of sleep. For the user to view this data, these wearable devices usually use a bluetooth radio to pair with the owner's smartphone. This bluetooth radio, however, allows for communication not just with the owner's smartphone, but also with the owner's environment. Since efficient and robust algorithms exist for determining whether the device is worn in a given instant, these platforms pose the potential for a yet unrealized new paradigm for the user to monitor not just their body but also their environment through communication with ambient bluetooth sensors. We believe this paradigm sets the stage for the next phase of home automation where the home is able to provide functions individualized to particular users such as TVs turning on to particular settings, shared vehicles moving to preset user settings, and home automation for particular habits. It also pushes the current paradigm of personal tracking forward by enabling tracking of habits which are exhibited not only by bodily signals such as step count, but also by environmental signals such as how long a wearer spends in different rooms of the house, uses different appliances, or spends in a car.

In the past, those with a compelling idea for a personalized computing application have been faced with a difficulty challenge. They had to bring together diverse skill sets ranging from hardware expertise to build the physical device, web or mobile expertise to create the interface, and machine learning expertise to provide the analysis. The difficulty of bringing together all these skills on a low budget made the success of new endeavors extremely unlikely as evinced by the recent failings of startup companies like Lively and Ninja Blocks. With the recent deployment of Android Wear, we aim to provide the next logical step. We develop a platform for rapid prototyping of personalized computing by pulling together the basic hardware, web/mobile, and machine learning building blocks. The core of this platform are three (mostly) open-source projects: 1) Android Wear, 2) TI Sensortag, and 3) Sci-Kit Learn. Using these projects requires only purchasing commercial off-the-shelf components such as an Android watch and TI Sensortag. Our project is built with generality, extensibility, and scalability in mind where methods are developed for hosting a full-stack application complete with extensive development both for computation on a local Android host and on a remote server.

In this work, we use the Sony Smartwatch 3, currently available for \$130, and the Bluetooth Smart TI Sensortag, currently available for \$30. We present results for one use case: home monitoring of an individual with dementia. After 12 months of beta testing, this use case is currently being deployed through the Dementia Care Ecosystem, a \$10M project sponsored for the Centers for Medicare and Medicaid Services. The rest of the paper continues as follows. Section 2.3 gives background on Alzheimer's disease and the specific use case. Section 2.4 describes the system architecture in detail. Section 2.5 describes the biggest challenges faced in development. Section 2.6 defines the analysis methods available and the subset included for the Alzheimer's use case. Section 2.7 describes a new indoor positioning

infrastructure for low-cost, room-level indoor positioning. Section 2.8 presents the results from beta testing. Section 2.9 provides concluding remarks and recommendations for future use.

2.3 The Dementia Care Ecosystem

One example new program is the Dementia Care Ecosystem sponsored by the Centers for Medicare and Medicaid Services [56]. The Dementia Care Ecosystem is a 3-year clinical trial evaluating a care model called Navigated Care, for people with dementia and their family caregivers. The goal is to improve quality of life, health care utilization, caregiver burden, and satisfaction with care. Central to the Care Ecosystem are minimally trained staff called ‘care team navigators’ (CTNs). These staff members are the primary point of contact for up to 80 families, allowing for personalized communication between families and their medical network. The Dementia Care Ecosystem is composed of four modules:

1. The **Caregiver Module** includes educational forums, caregiver support, and connects families with community resources.
2. The **Decision-Making Module** facilitates proactive medical, financial, and safety decisions.
3. The **Medication Module** tracks and reduces inappropriate medications or doses, and triggers pharmacist review when indicated.
4. The **Functional Monitoring Module** uses smartphones and sensors to rapidly detect and respond to changes in functional status, which is particularly important for patients living remotely, alone, or who are at-risk for acute declines.

The use case of the system described in this paper is for the functional monitoring module. The goal of the functional monitoring module is to calculate five metrics and provide alerts when these metrics deviate from expected. These five metrics include daily step count, approximate gait speed, daily lifespace, daily and hourly room percentage, and daily room transitions. Based on these five metrics, 2 sets of analyses are conducted: outlier detection and trend detection. Data is collected with an Android smartphone, Android smartwatch, and in-home sensors, analyzed using a backend server, and displayed for CTNs using a Salesforce dashboard. CTNs respond to alerts by confirming the data appears abnormal then calling the family or clinical team as needed.

This system is the first long-term, continuous, personalized monitoring system developed for individuals with cognitive disorders. This represents part of a growing paradigm shift from medical equipment that is designed to react to disease states (e.g., X-ray imaging), to equipment which is designed to enable a proactive medical system. It is the first long-term. Many other technology systems for Alzheimer’s care are available including offerings from companies like HealthSense and BeClose, but they fail to identify a person uniquely.

Thus, the primary impetus for this system was the need for a system which could monitor the behavior patterns of a specific individual in the presence of multiple individuals living in the same residence. The greatest challenge in accomplishing this goal was designing a system capable of robust indoor positioning using cost-effective off-the-shelf equipment. The solution is described in Section 2.7.

2.4 System Architecture

The goal of the system architecture is to provide a pipeline whereby data from the external environment can be collected, analyzed, and used for generating notifications if necessary. For the Alzheimer’s monitoring use case, the pipeline is used to calculate five metrics, determine possible causes for alarm, and raise notifications for care team navigators responsible for monitoring affected individuals. The full system architecture instantiated for this use case is shown in Figure 2.1.

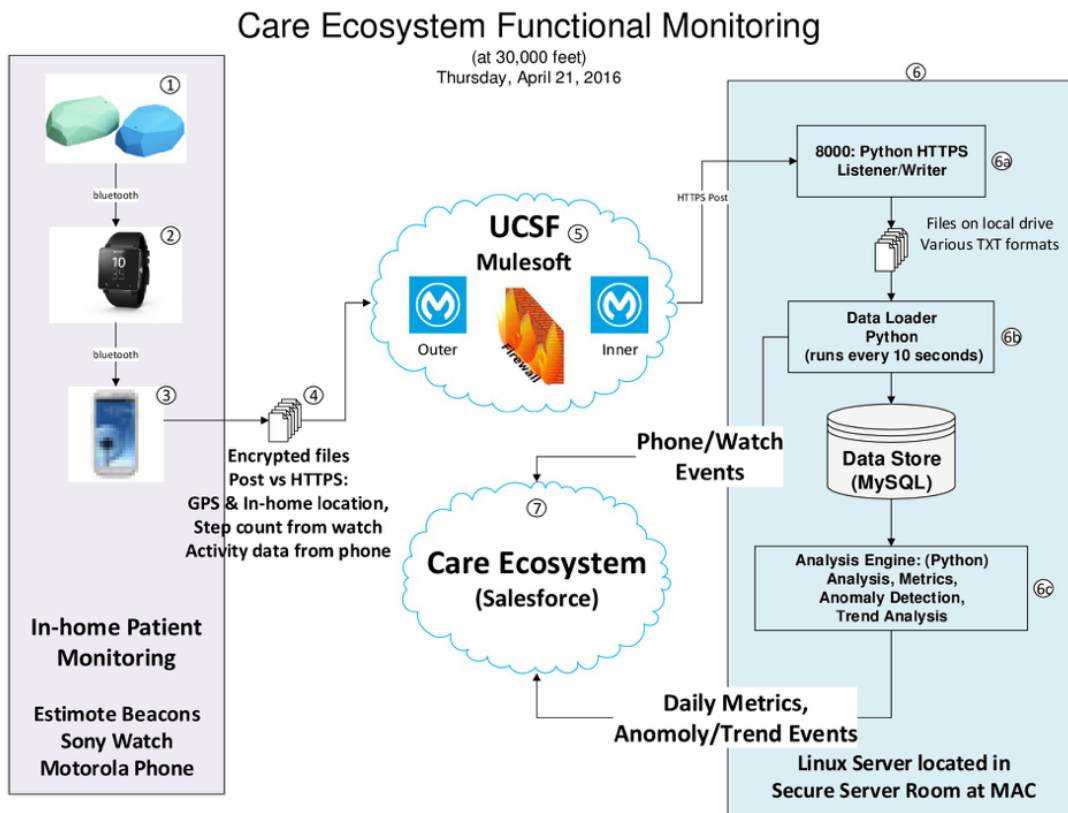


Figure 2.1: The architecture deployed for the Dementia Care Ecosystem.

The five metrics used for the Dementia Care Ecosystem include:

1. **Step count:** The number of steps taken per day
2. **Approximate gait speed:** The steps per minute during each period of activity
3. **Lifespace:** The maximum euclidean distance from the home per day
4. **Room percentage:** The percentage of time spent in each room on a daily and hourly basis
5. **Room transitions:** The number of transitions between rooms per day

Although these are the five metrics chosen for the Dementia Care Ecosystem, we take care to design a system which is flexible enough to handle the inclusion of any Bluetooth sensors. Off-the-shelf, the application allows for the collection of all data sources from the smartphone and smartwatch depending on the capabilities of each. For instance, although we do not use heart rate data, it can be collected at the desired interval simply by connecting a smartwatch with a heart rate sensor to the smartphone through standard Android Wear procedures then updating the appropriate settings.

To collect these metrics, the system architecture is composed of five units. The Estimote beacons output a Bluetooth low energy (BLE) ping at a user-defined power and frequency. The Android smartwatch acts as a BLE receiver for in-home sensors, collects step-count data, and buffers data until it can be transmitted to the smartphone. The Android smartphone collects GPS data, provides a user interface (UI), and uploads data to the backend server at fixed 12-hour intervals. The Android smartphone communicates with the backend server through an optional security proxy called Mulesoft. The backend server performs data processing to estimate the room location, calculate desired metrics, flag outliers, and detect undesirable trends. The Salesforce dashboard provides an interface for administrators to receive alerts and view metrics from the backend server. The decision to push most of the computation to the server was made to preserve battery life. This allows for prototyping before isolating those functions which must be performed in real time. For instance, room estimation is performed offline with the method that provides the highest accuracy for our use case, but can be performed locally using the methods discussed in Section below.

BLE Sensors

Two types of Bluetooth low energy (BLE) sensors are used: one to provide consistent room locationing and another to infer activities of daily living (ADLs). The sensors used for room locationing are called Estimote beacons. They provide a simple BLE ping with customizable power settings, so the receiving device may output a received signal strength indication (RSSI) which is roughly proportional to the distance from the beacon. The second type of sensor used is the TI Sensortag. Equipped with 10 different sensors including temperature, motion, and humidity, the Sensortags allow for collecting many types of data which is useful for inferring the performance of different activities (e.g., medicating). These two sensor types

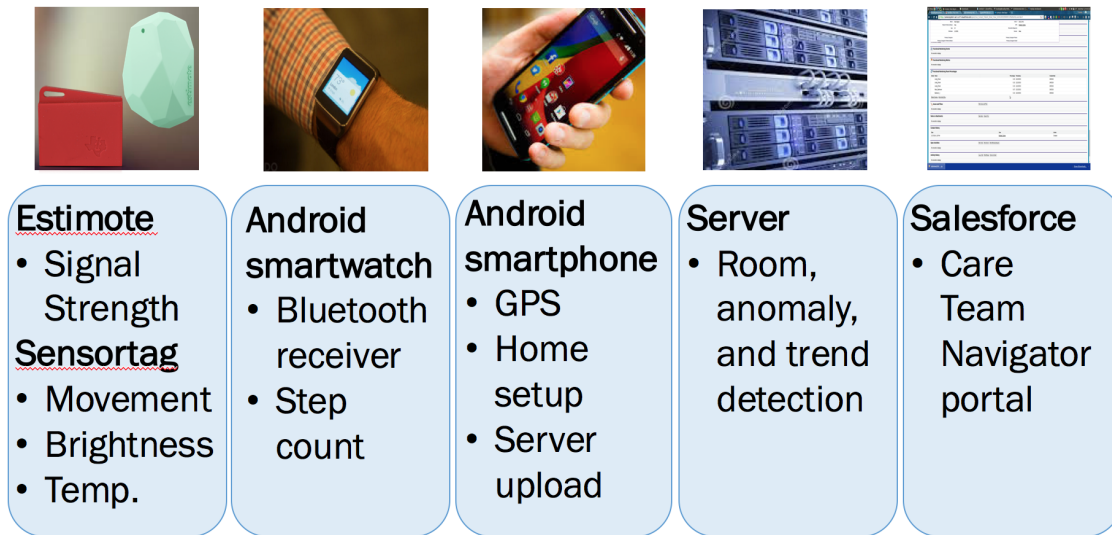


Figure 2.2: The five system components and their uses.

are actually redundant. The Estimotes are equipped with temperature and motion sensors, and the TI Sensortags provide RSSI values. In practice, using both sensors allows for rapid prototyping which is both flexible and modular where users have the option to only include those sensors most relevant for their use case.

The sensors used to perform room detection are called Estimote beacons. The Estimote beacons provide a simple BLE ping to infer distance from the Android application. The rate at which the beacons ping is set to 200ms by default and the smartwatch scans for these pings every 10 second for a 1 second period by default. The Estimote beacons with sticky backs can be placed on ceilings or walls without reducing the aesthetics of the room.

The sensors used for inferring user activity are called TI Sensortags. They provide sensing for 1) temperature, 2) acceleration, 3) orientation, 4) humidity, 5) magnetic flux, 6) ambient light, 7) pressure, and 8) audio. Because the Sensortag is equipped with so many types of sensors, it allows for prototyping of many different features with minimal changes to source code. Data is sampled from the Sensortags as follows. Every minute the watch scans for available tags. It then connects to the three tags with the highest RSSI value. This gives some measure of which tags are closest and likely to provide data related to the user activity. Following connection, data is streamed at the desired interval from the desired sensors on the tag. For example, tags identified for monitoring shower user provide readings from the humidity and temperature sensors at a slower rate than tags identified for identifying if a specific object has moved via the accelerometer. Because data from the Sensortags can quickly grow to large scales, this data is stored in as bson (binary json) and only transmitted to the phone once per hour to conserve battery by default.

Thus, Max allows collecting data and testing algorithms for different functions, but once battery life becomes a constraint new methods may need to be devised. Because the Sen-

sortag software is open source, the user can implement simple detection algorithms such as the spike events discussed below on the tags themselves and only transmit resulting detection to the watch. This would allow for Bluetooth beaconing mode instead of pairing mode to be used, reducing battery consumption from both the sensor and the watch.

Smartwatch

The smartwatch used to determine the wearer's location in the home and to collect stepcount data is the Sony Smartwatch 3. We choose this watch because it provides a longer battery life than the Moto360 or Samsung Gear Live (no longer produced) based on empirical tests. It further provides IP68 dust and water resistance, dust tight and submersible in water up to 1.5 meters. It should be noted that this water resistance requires a charging cap to be closed which cannot be guaranteed in the use case discussed here. As discussed below, the battery on the smartwatch is the greatest bottleneck in the project for which reason extreme care has been taken to develop efficient asynchronous data collection from the smartwatch. Once data is collected, it is stored in a sqlite database on the watch and uploaded every hour to the phone by default.

Smartphone

The smartphone used to collect GPS data, provide a user-interface, and upload data to the backend server is the Motorola Moto G. This phone was chosen because it was the most cost-effective phone supporting the latest version of Android. At the time of publication it cost \$150 retail. The system developed is not phone specific, however, and has been tested on a number of Android phones. The smartphone uploads data to the backend server by default every 12 hours. GPS data is recorded whenever a change greater than a certain threshold occurs through standard Android protocols.

Server

The server used to perform indoor positioning, outlier detection, and trend detection is a high-performance desktop computer located at the UCSF Memory and Aging Center. The hardware specifications include 32 GB of RAM and 1 Intel Quadcore x86 CPU operating at 2.7 Ghz. The operating system used is Linux Ubuntu 14.04. The server hardware specifications can be chosen to match the algorithms and scale required. The server operates as follows. An http receiver asynchronously uploads data to a SQL table responsible for storing all metrics. When new data is received, functions are applied to calculate the required metrics (e.g., percentage time spent in each room). After metrics are calculated, analysis is performed to detect anomalies and trends. If any detection occurs, events are sent through the Salesforce API.

Dashboard

The dashboard for the administrator to observe events detected is implemented using Salesforce. For this use case, events are detected on a daily basis based on the data from the previous day, so the care team navigator can review adverse situations and handle them as needed. The goal here is for the clinical team to collect the data necessary to determine which information is most useful for evaluation before translating the system into a device most suitable for patient and clinician needs.

2.5 Biggest Development Challenges

Full system development from initial prototyping to deployment with patients required 18 months from 3 active developers. The greatest challenges involved battery life, always-on functionality, security, robustness, and methods to encourage adherence specific to the use case.

Battery Life

Battery life concerns limited the amount of data that could be collected and increased the latency between data collection and analysis. We strove to maintain 1-day battery life to avoid any deviation from normal routines which many users would find overly burdensome. In order to accomplish this, we found that sampling the sensors on the watch caused the greatest battery drain. As shown in Figure 2.3, this prevented the CPU from sleeping properly and led to rapid battery loss. We explored storing data in a hardware FIFO [30], but found the best solution to be focusing in on what data we really needed. We read the step count directly from a chip present in many smartwatches where the step count algorithms is implemented directly in hardware. We read gyroscope data once per minute, infrequently enough to prevent excessive battery drain. We then calculate the variance of this data over a rolling 30-minute window and see if it has passed an empirically defined threshold to determine if the watch is currently on the body or not. We further conserve battery life by reducing the sampling rate from the Estimote beacons when the user is determined to be outside the house by GPS or because an Estimote ping has not been received in a preset amount of time.

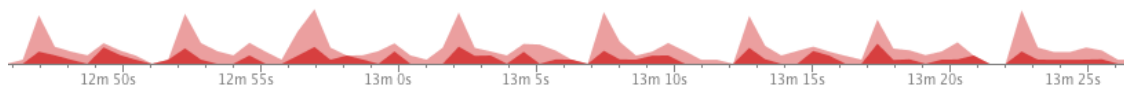


Figure 2.3: CPU usage before optimizing for battery life. Waking up the CPU frequently to sample from the sensors caused rapid battery loss.

Always-on Functionality

In order to maximize data collection time, the Android application starts itself when the phone turns on and cannot be closed by the user. In this way, data collection is ensured unless either the phone or watch is turned off. This always-on functionality is accomplished through the use of the notification services provided by Android (foreground service) that are independent of the application user interface. Thus, unless the user forbids the notifications, there is no way to stop the service, but the user is always informed by a notification when monitoring is underway. As a further failsafe, a background service independent of the others regularly checks if the other services are running, and if not, restarts them. This keeps services running even if the application crashes.

Security

Before deployment with patients, the platform was required to pass security review at UCSF. The end result includes four measures to increase security.

Encryption

Data is encrypted in the phone and watch internal sqlite database. When it is transmitted, it is done through a secure https tunnel and decrypted on the server.

Mulesoft

Mulesoft is used as a security proxy following standard protocol at UCSF, preventing the need for the Android application to store a private server key. Mulesoft further provides a simple mechanism for archiving all encrypted data before it reaches the backend server. Passing the Mulesoft security review at UCSF required installing a CA certificate on the backend server; using HTTPS for pushing data from Android to Mulesoft and from Mulesoft to the backend server; creating Maven build profiles for each of the development, staging, and production environments; exporting all server URL's, domain names, ports, and related networking information to external property files; and modifying error logging to only record the most severe errors.

HockeyApp

HockeyApp is used for private distribution of the Android application. It allows sharing the Android application with new devices via email and pushing software updates without requiring posting the application publicly on Google Play. HockeyApp can be used for free with up to 2 apps and unlimited storage, crash reports, users, and user feedback. We also explored creating a Google Play private channel, but found the \$50/user/year fee to be excessive. The most challenging part of incorporating HockeyApp was implementing automatic updates. By default HockeyApp only checks for updates when the app is opened.

For the Dementia Care Ecosystem use case, however, it was expected that the app would operate continuously in the background without the user ever necessarily opening the app after the initial home installation. Unfortunately, HockeyApp provides no API for updating the application on a regular basis, so automatic updating was implemented by decompiling their library and reimplementing automatic checks for an update in a background service responsible which is active whenever the phone is on through the always-on functionality described previously.

Progaurd

Progaurd provides Android enabled code obfuscation. We make it available in this system, but disable it by default since we found it caused too many dependencies to be broken within our code. It also only makes it slightly more difficult for a would-be attacker to find a password or ssh credentials hidden within the code since the password itself will still be present in the code even if the variable name is obfuscated. We thus prefer the Mulesoft security proxy discussed above to any code obfuscation techniques where security is required.

Robustness

To ensure robustness, the system was beta tested by 13 different healthy subjects over 39 total months months before deploying with real individuals affected by Alzheimer’s disease. This testing exposed unexpected challenges including difficulty with certain ceiling types when attaching the Estimote sensors, difficulty connecting to the backend server through certain routers, difficulty connecting to certain beacons, and many other smaller issues. This led to numerous bug fixes and an extensive trouble shooting guide for new installations. After such extensive testing, we are happy to share Max with such high confidence in its robustness.

Adherence

To maximize adherence, we detect when the watch has not been worn, when the watch-phone connection has been lost, and when either the watch or phone is out of batteries. We alert the caregiver through email or text message based on personal preference. The preference is determined during the initial home setup. The on-body detection is performed with the gyroscope as described above. The break in watch-phone connection and out-of-battery signal are both provided by the Android API.

2.6 Analysis Methods

Two sets of analyses are provided to determine possible causes for alarm based on metrics collected.

1. **Outlier Detection:** Two methods for outlier detection are provided, DBSCAN and a nearest neighbors approach from [29]. These methods are used to determine if any metric significantly differs from a baseline which is set to 30 days by default. Both are used in a fully unsupervised setting.
2. **Trend Detection:** The RANSAC algorithm is used to determine if any metric has declined by greater than a certain threshold over the a previous window. In the dementia use case, we examine if any metric has declined over 33% over the previous 30 days.

DBSCAN

DBSCAN performs clustering based on proximity. The DBSCAN algorithm or Density-Based Spatial Clustering of Applications with Noise algorithm clusters points using two hyperparameters, the minimum number of points required to form a cluster and the maximum distance two points can be apart to be considered within the same cluster. The clusters are formed by choosing an arbitrary starting point, connecting all points in the cluster, then choosing an arbitrary new starting point. The process is repeated until all points are either assigned to a cluster or identified as not belonging to any cluster in which case they are labeled outliers.

This method of outlier detection is provided because it provides good results in many use cases and allows for seamless substitution with other methods from the Sci-Kit Learn API where many other outlier detection methods are available. We prefer the method discussed next which was implemented for this project due to its higher level of support in the outlier detection literature.

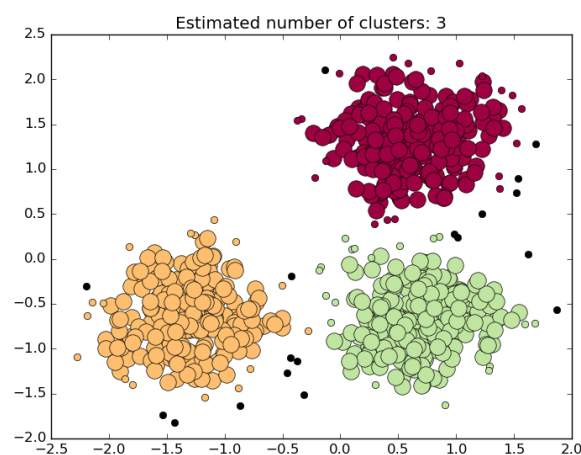


Figure 2.4: Outliers are those points labeled by DBSCAN as not belonging to any cluster [53].

k-NN Outlier Detection

As recommended by [29], we prefer *k*-NN for global outlier detection tasks. This method has been shown to provide high performance on a diverse applications ranging from breast cancer diagnosis to handwritten digit classification. The caveat is that these techniques work well in low-dimensional spaces where distance metrics to determine the nearest neighbors can be calculated in a manner which is meaningful and efficient. In all applications discussed in [29] where *k*-NN produced the best or very good results, *k*-NN was applied in a low dimensional feature space ($d \leq 40$).

Outlier detection with *k*-NN is performed as follows. For each point, an outlier score is calculated by choosing the *k* closest points and determining the average distance to these points. We choose $k = 4$ and the Euclidean distance metric by default. This outlier score is used to define an ordering over the points. With this ordering defined, the most anomalous point is that with the greatest outlier score. It is the point furthest from its closest neighbors. After defining this ordering, some threshold must be set to determine what values constitute an outlier. In the semi-supervised setting, this threshold may be determined empirically from data or from thresholds which may be relevant for the use case. Before data collection begins, however, there may be no way of determining an appropriate threshold. For this case, we provide thresholds based on percentiles. By default, a severity level 1 event is triggered when it is in the top 3% of the ordering, a level 2 event in the top 1%, a level 3 event in the top 0.3%, a level 4 event in the top 0.1%, and a level 5 event in the top 0.03%. For the metrics defined in the dementia use case which are measured on a daily basis, a level 1 event will occur in normal data approximately every 30 days and a level 5 event will occur in normal data approximately every 3000 days – much like using the term 100-year storm to describe a weather event so severe it should only happen once every 100 years.

The challenge with these approaches is that they view the data as stationary. As shown in Figure 2.5, as the data distribution changes over time as would be expected in the dementia use case, these methods will fail to identify the shift and instead identify those points which are furthest from the baseline as anomalous. If in actuality the distribution is shifting, these points may signal an important trend rather than an anomaly, but will still be highlighted as anomalous until enough of them occur. In situations where the stationarity assumption is violated, we recommend detecting anomalies based on a rolling window where the severity levels are determined not for the whole data set, but for windows of various sizes which can be thresholded in the same manner as discussed above.

RANSAC

The default method for trend detection is RANSAC [20] linear regression. The RANSAC algorithm or RANdom SAMple Consensus Algorithm performs model fitting in the presence of outliers. In the original method, hyperparameters defining the tolerance and threshold are provided [20]. In the method defined by sci-kit learn [53] and used here, the maximum number of trials is provided. The algorithm iterates by selecting a random sample of the

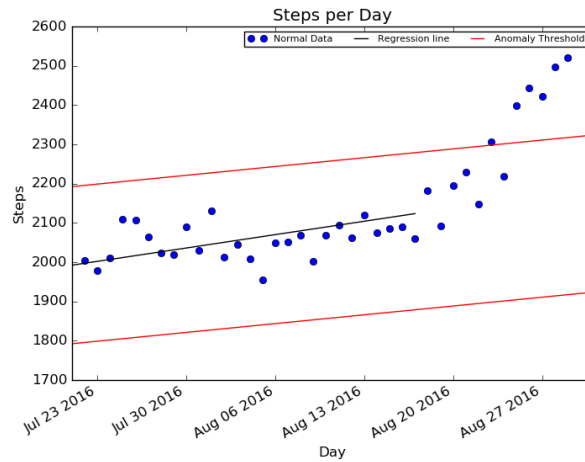


Figure 2.5: Global outlier detection with k -NN fails for a nonstationary distribution. Local outlier detection methods provide empirically worse performance as described in [29]. We instead prefer the k -NN with a rolling window when anomaly detection over a nonstationary window is required.

data containing the minimum number of points required to estimate the model parameters. In the case of linear regression with 2-dimensional data (e.g., daily step count), this requires 2 points. From this sample, a model is fit. By [20], if the predefined threshold of points fit within the tolerance, the algorithm terminates. If not, the program iterates. In practice, this method can result in infinite iteration if no acceptable model with the predefined hyperparameters exist. [53] instead simply repeats for a predefined number of iterations and then chooses the parameters which result in the greatest number of inliers. This surprisingly simple and highly computational efficient model grew out of the computer vision community and has found wide success on a number of applications for which reason it is the default here. Note that alternative methods such as the Theil-Sen regressor for median fitting are provided by Sci-Kit Learn and can be easily substituted given the matching API.

2.7 Indoor Positioning

As discussed in Section 2.4, in order to perform indoor positioning, Estimote beacons are used and the received signal strength indicator (RSSI) is detected by the smartwatch. The RSSI is filtered then the approximate location is determined through supervised learning. This method allows for approximate positioning to be performed by matching the wearer of the watch to the closest beacon. It was chosen over methods based on triangulation that require multiple beacons per room to ensure at least 3 beacons are visible at all times. In contrast, this method allows for cost-effective approximate indoor positioning by binning users into one location from a set of possible locations. In the Care Ecosystem, it is used

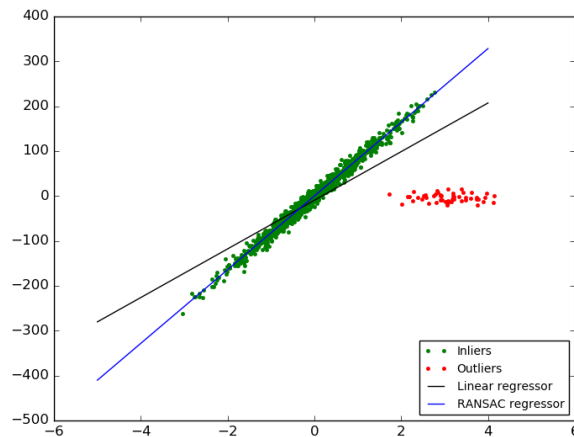


Figure 2.6: RANSAC fits a regression line by consensus, providing robustness to noise [53].

for cost-effective room-level indoor positioning where one beacon is placed in each room. By default, the RSSI values are adaptively filtered with the kernel least mean square (KLMS) algorithm then labeled with the random forest supervised learning algorithm; however, many more techniques for filtering and supervised learning are provided. Both default methods are empirically chosen based on results demonstrated in the following section. It should be noted although we do not provide off-the-shelf support for triangulation, the same Estimote beacons can be used for this function if needed.

Radio Wave Propagation and WPL

The primary challenge in accurate indoor positioning is the highly variable nature of the radio wave received signal strength indicator (RSSI). Because standard construction materials provide moderate impedance to radio waves, signals may be both transmitted and reflected through the surrounding environment. Thus, certain regions of the room may demonstrate high RSSI despite being further away from the source. Moreover, changes in the room environment such as people walking to different positions significantly alter these multipath effects. For this reason, fingerprinting techniques that attempt to laboriously map the RSSI of the room struggle to maintain robustness [52]. To compensate, these methods often use techniques like k-nearest neighbors (kNN) to increase accuracy, but show only marginal gains within an individual block of signal strength emitters (e.g., when only one emitter is used per room). In order to account for the exponential decay of radiowaves, many nonlinear filters have been applied including particle filters, extended Kalman filters, path loss models [52, 82, 76]. We leverage the path loss method here because it provides a natural fit as a feature in supervised classification. The path loss model used is based on the ITU Indoor Propagation Model [82] in which the signal strength can be expressed as the path loss over

a distance d (m) at frequency f (mHz)

$$PL(d, f) = 20\log(f) + 10\alpha\log(d) + c(k, f) - 28 \quad (2.1)$$

where α is the path loss exponent, k is the number of floors between the transmitter and receiver, c is an empirical floor penetration loss factor, and f is the radio frequency. With f considered constant in this case for Bluetooth at 2.4 GHz, the signal strength can be expressed as

$$PL(d) = PL_0 + 10\alpha\log(d) \quad (2.2)$$

In the weighted path loss model (WPL), the indoor propagation model is used to estimate position based on the RSS [82]. Weights are assigned by solving equation (2) for d and defining the weighted factor for the i th RSSI as

$$w_i = \frac{1/d_i}{\sum_i 1/d_i} \quad (2.3)$$

The unknown position of the person is then estimated as

$$(q, r, s) = \sum_i w_i(x_i, y_i, z_i) \quad (2.4)$$

where (x_i, y_i, z_i) is the position of the i th beacon. WPL has traditionally been used to replace techniques like kNN as a supplement to fingerprinting. We use it here to define the kernel for KLMS.

KLMS

Adaptive filtering techniques provide a framework for estimating a non-stationary signal. They converge to the optimal linear filter in the mean square error. The Kernel Least Mean Square (KLMS) algorithm is a technique for adaptively filtering nonlinear data.

As an adaptive filtering technique, KLMS requires an iterative convex optimization algorithm to converge to the minimum mean square error. KLMS traditionally is the application of just one technique, the popular stochastic gradient method, but many convex algorithms can be applied to affect the convergence of the filter. We prefer stochastic gradient descent with Nesterov momentum here due to the increased convergence rate. The general equation for gradient descent with Nesterov momentum is:

$$x_{n+1} = x_n - \mu\nabla f(x_n + \beta(x_n - x_{n-1})) + \beta(x_n - x_{n-1}) \quad (2.5)$$

where $0 < \beta < 1$ defines the momentum hyperparameter. Note that if $\beta = 0$, no momentum is present and the iterates are the same as the gradient method. Compared to traditional gradient descent, this method is more robust to ill-conditioning and provides faster convergence bounds which affects how well the adaptive filter approaches optimality.

In stochastic methods, rather than accessing the gradient directly, we compute a function with the same expected value [61]. That is, we can approximate the gradient by a function $g(x)$ such that $E[g(x)] = f(x_n)$. In KLMS, this amounts to minimizing the mean square error over a fixed number of filter taps rather than the true mean square error. As in other stochastic methods, because we replace the actual function by one that only shares the expected value, we now converge only in the expected value. That is, some randomness will be introduced into our convergence and we will converge to a ball with some radius rather than a fixed point.

The end result is a minor update to the traditional KLMS method to perform stochastic gradient descent with Nesterov momentum where

$$w_{n+1} = w_n + \mu e_n x_n^H \quad (2.6)$$

is updated to

$$w_{n-1} = w_n + \mu(d_n - (w_n + \beta(w_n - w_{n-1}))^T x_n) x_n^H + \beta(w_n - w_{n-1}) \quad (2.7)$$

As with traditional KLMS, this update rule is applied in the kernel space not the time domain. This KLMS with momentum is applied to the RSSI values from each beacon independently as a univariate analysis. The result is that correlations in the time domain are handled through filtering, so that the next supervised learning phase can handle each point as if it is independent.

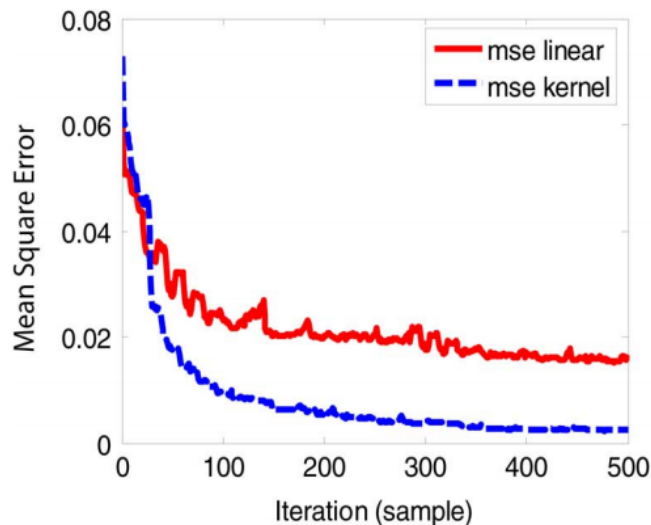


Figure 2.7: KLMS outperforms linear LMS in fitting nonlinear functions when the nonlinear function class is known a priori [45].

Random Forest

Many supervised learning techniques are available through Sci-Kit Learn. In our room-estimation pipeline we try several methods and choose the one which produces the best result over 3 fold cross-validation. The best resulting classifier is often the random forest classifier. We thus describe the random forest classifier here to provide full insight into one potential pipeline with the caveat that the Sci-Kit Learn API is nearly identical for all supervised learning methods, so any number of alternate methods are available for substitution.

The random forest classifier and related methods such as extra trees, gradient boosting and AdaBoost increase accuracy by ensembling many weak classifiers, a technique known as bagging or boosting depending on how the ensemble is formed and accumulated. The weak classifier for the random forest is the decision tree, and the random forest is formed by averaging over the predictions of many decision trees. The difference between the two can be seen in Figure 2.8. The resulting classifier is improved if each of the weak classifiers is similarly good, but uses different features to form the decision boundary (i.e., averaging over many instances of the same decision tree will produce a result no better than the original decision tree). For this reason, randomness is injected by selecting a random subset of the available features and forming the best decision tree from this subset.

The fact that the random forest performs well in this scenario highlights that even after filtering with a nonlinear kernel, the resulting data points remain difficult to separate with a linear decision boundary. This suggests the RSSI data is highly nonlinear not only due to the exponential decay of radio waves, but also due to the variable impedances present in the environment.

2.8 Results of Beta Test

We present results from 39 total months beta testing the features of Max required for the Dementia Care Ecosystem use case.

User Interface

The results are available for viewing through Salesforce. Max provides the methods necessary to transmit metrics and analysis to Salesforce through a predefined API. The implementation in Salesforce itself was performed by contractors at UCSF with the end result shown in Figure 2.9.

The primary function of the Android user interface is to allow new users to perform initial home setup and allow administrators to view appropriate debugging information. Examples of each are shown in Figure 2.10. This Android user interface can be extended and customized through standard Android development techniques. As a sidenote, when developing user interfaces in collaboration with individuals without an engineering background, we found InVision to be an extremely useful service. InVision allows users to develop the look and feel of an application in PowerPoint, a more broadly available skill.

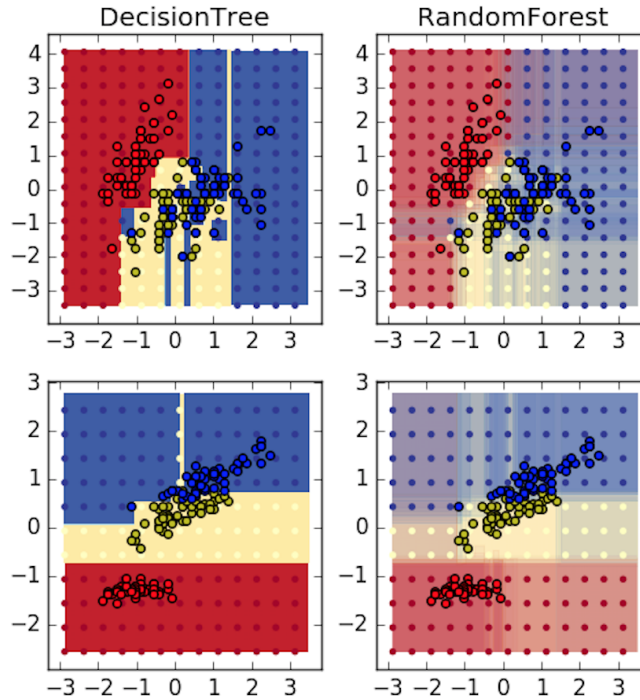


Figure 2.8: The random forest increases classifier accuracy by averaging over random decision trees to reduce variance [53].

Room Estimation

Based on 13 home setups with number of rooms equal to 3.1 ± 1.7 , the room detection accuracy is $96.1\% \pm 2.6\%$. In these procedures, only one was performed in a house with rooms on multiple floors. More rooms and more floors would naturally decrease the detection accuracy as the number of neighboring rooms increases. Room sizes were allowed to vary as they naturally do in the home setting with small rooms on the order of 1 meter diameter (e.g., bathrooms) and large rooms on the order of 6 meter diameter (e.g., living rooms). Estimate beacon settings were all set to the same parameters with broadcasting power set to -20 dBm, sufficiently large to cover any room size, and advertising interval set to 200 ms, sufficiently small to provide many opportunities for detection even when just passing through a room.

Results from one representative plot are shown in Figure 2.11. The top line of circles shows the true room at each point in time. The line below it of triangles shows the prediction made at each time point. The squares scattered below show the RSSI value from each beacon at each point with higher values denoting beacons expected to be closer. As shown, the method is able to successfully resolve uncertainty when the RSSI value is fluctuating between two rooms, a situation which is highly detrimental to the Care Ecosystem use case in which a key metric is the number of transitions between rooms. The cost is decreased accuracy

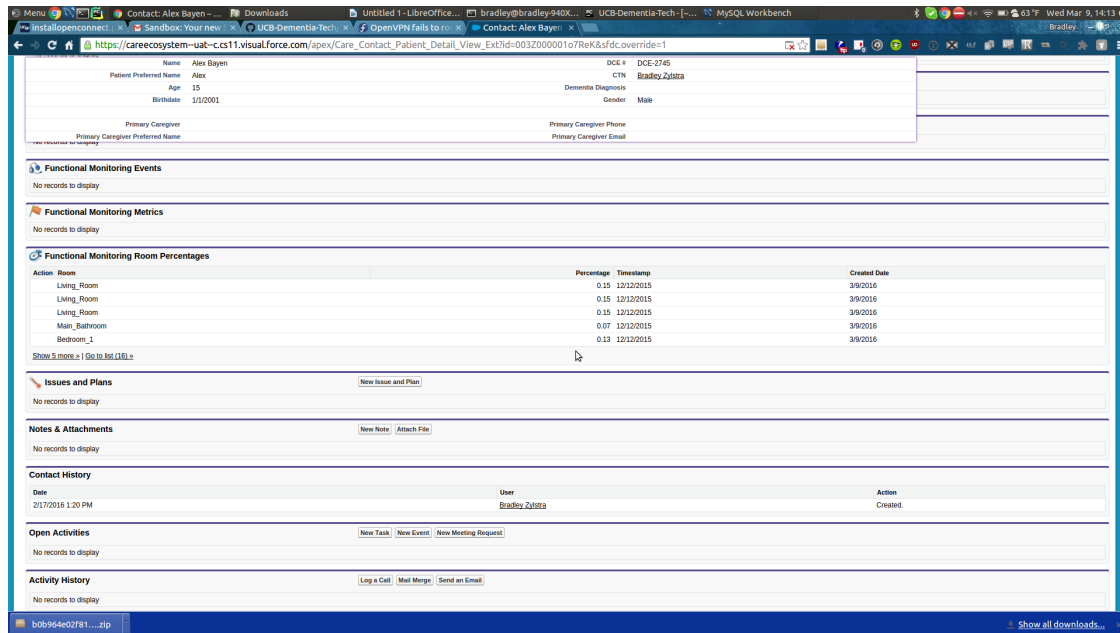


Figure 2.9: Salesforce user interface for care team navigators to view metrics and analysis for monitored individuals with dementia

when a true room transition is made. In most situations this occurs far less frequently.

Analysis

After data is collected from the watch and ambient sensors, metrics can be formed and analysis methods applied. Two example results of the analysis are displayed in Figures 2.12 and 2.13. In Figure 2.12, outlier detection is applied after room estimation has been applied to infer the percentage of the day the user spends in the Bathroom and the Bedroom. In this situation, the home environment was an apartment limited to these two rooms. From the outlier detection, a distinct pattern emerges. On most days, the user spends very little time in the bathroom. On some days, the user spends more time in the bathroom. This cluster was mostly composed of days in which the user showered. Finally, an outlier is detected when an atypical point was detected from the normal. In the Dementia Care Ecosystem, this would flag the care team navigator to call the user and ask about specific symptoms defined by a flowchart designed by the clinicians involved. In Figure 2.13, trend detection is applied to the step count data collected on the watch in the first 30 days of use. The use of the fitness tracker appears to show the desired increase in number of steps taken over this time. Note that the two points in which many steps are taken are ignored in the resulting model. Similarly, this robust RANSAC regression allows model fitting in the presence of many days when the user forgets or chooses not to wear the device.

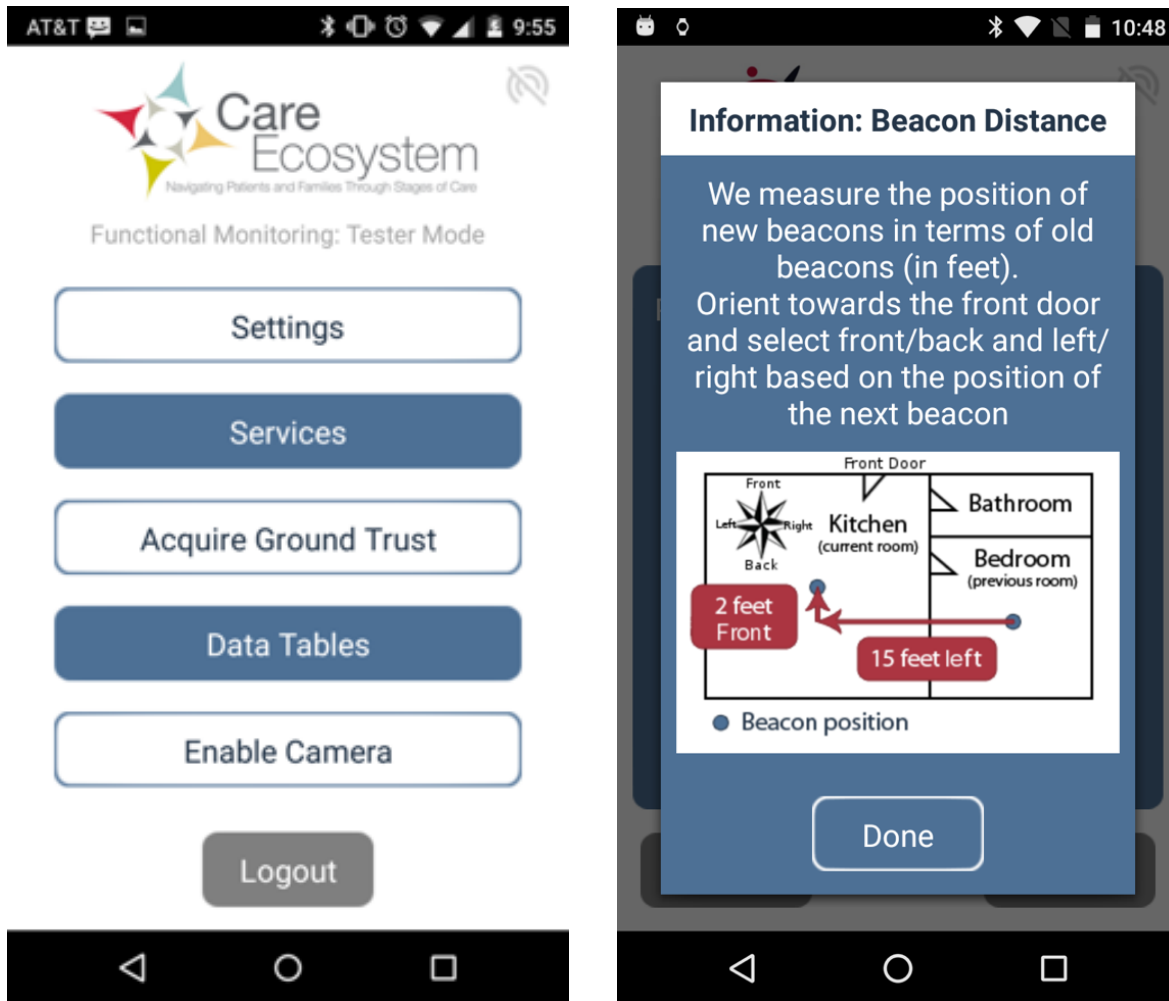


Figure 2.10: The Android user interface for administrators and a selected screen from home setup. The interface for non-administrator users provides a subset of the functions available.

2.9 Conclusion

In this work, we describe Max, an open source prototyping platform constructed from off-the-shelf components for designing cyber-physical systems with personalized to individual users. The current starting price is \$400 assuming the smartphone and smartwatch used here, three sensors, and an available server for computation. At this price, many interesting new applications are feasible. We hope Max reduces the engineering burden of creating such systems to spur innovation in creating new and interesting applications. We describe one such application here in the Dementia Care Ecosystem. The Dementia Care Ecosystem aims to reduce the cost of dementia care through cost-effective continuous monitoring to detect changes in behavior early enough to respond before a painful and expensive emergency room visit may be needed. One example is monitoring changes in bathroom use for signs of possible

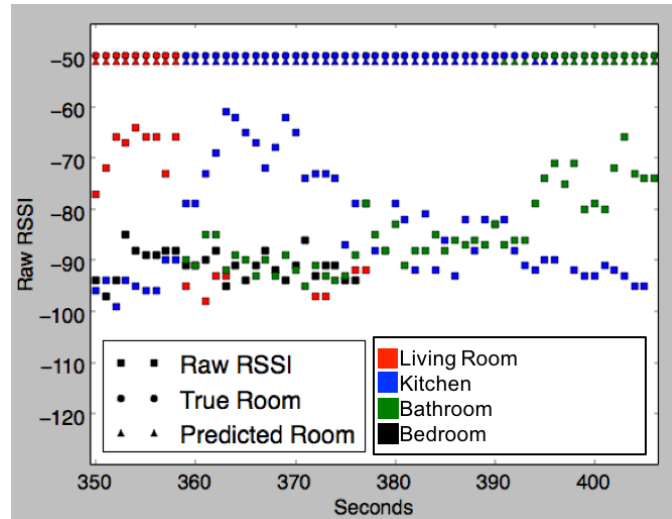


Figure 2.11: Representative plot of room inference from raw RSSI

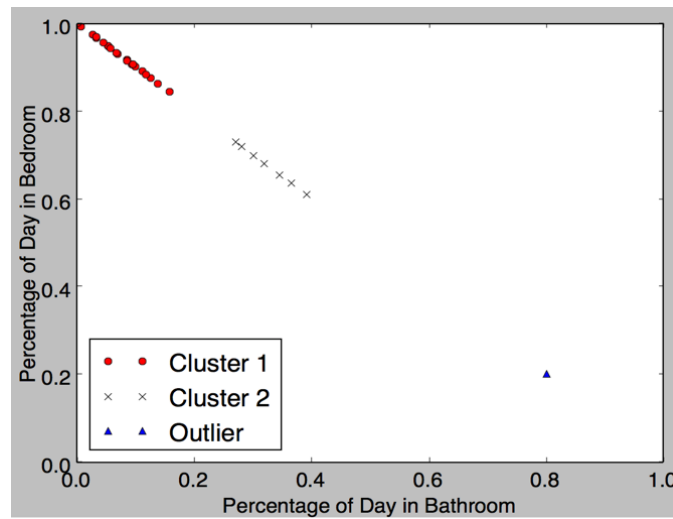


Figure 2.12: Representative plot of inferred room location with outlier detection applied

urinary tract infections, an unfortunately common problem in dementia care.

We present the system architecture for collecting data, maintaining data securely, and performing several common data analysis techniques including filtering, classification, anomaly detection, and trend detection. Where possible, we give concrete examples from the Dementia Care Ecosystem use case and highlight where other methods are available through open-source libraries such as Sci-Kit Learn. We further derive and demonstrate the efficacy of a new technique for cost-effective approximate indoor positioning, a common need for many personalized applications which is not met by current offerings.

There are several features that we plan for future inclusion in Max but are not yet avail-

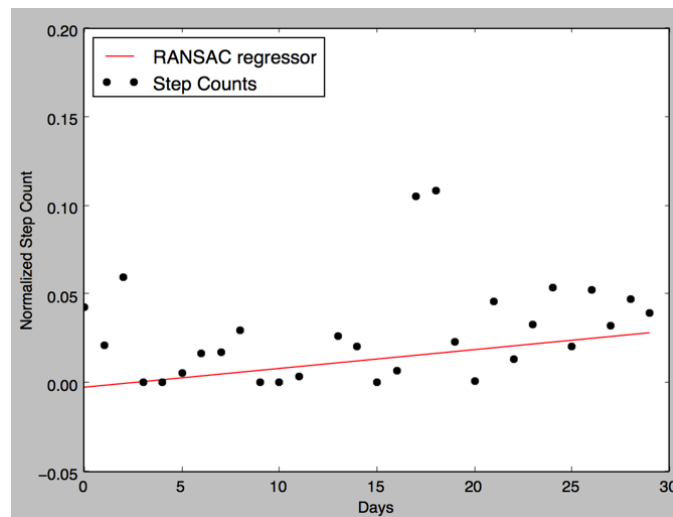


Figure 2.13: Representative plot of step count data with trend detection applied

able. One significant missing feature is the presence of any actuators. Given the current disjointed market for sensors and actuators, it is difficult to establish a common API for inclusion of the many types which would be interesting to use with Max. By building this project as an open-source collaboration, we hope to gain support from the community in developing support for prevalent IoT platforms. Some of the notable platforms for which we will encourage inclusion in the near future include Automatic for automotive and Samsung SmartThings for home. Another missing feature is the lack of analytical methods which leverage large quantities of data for increase performance. To this end, we anticipate the future inclusion of Lasagne and Theano, open source libraries for deep learning and computational graph analysis. For example, as the Dementia Care Ecosystem scales, this inclusion will enable the ability to label sequences using recurrent neural networks based on the annotated data already collected through the current implementation.

The challenge from a hardware perspective now stems not from the wearable device, but from the surrounding sensors. We thus conclude that if those who wish to drive innovation in the wearable computing market such as Google and Apple similar to that in current smartphone ecosystems, they should produce a developer's kit with the sensors and actuators necessary to enable a large array of potential applications. The seamless inclusion of these sensors would dramatically reduce the burden of producing the next generation of wearables, where interaction is enabled not only with the user's body but also with the surrounding environment.

Chapter 3

Diagnosis through Speech

3.1 Chapter Abstract

The aim of this work is to provide computer tools to help diagnose subjects with various dementias by applying machine learning algorithms to recorded conversations between patients and close caregivers. The dataset includes 126 conversations collected between 2002 and 2014 and including patients with Alzheimer’s Disease (AD), behavioral variant Frontotemporal Dementia (bvFTD), Primary Progressive Aphasia (PPA), and healthy controls (HC). By combining both acoustic and text features, we reach a level of 92% accuracy in distinguishing dementia from healthy controls and 75% in distinguishing between subtypes (AD vs. bvFTD vs. PPA vs. HC). Most notably, by collecting more than 1200 features and selecting the most relevant ones, we highlight highly relevant features that cannot practically be collected by a human during clinical observation, suggesting new avenues for computer-aided diagnosis and prognosis of dementia.

3.2 Introduction

In this chapter, we develop, prototype, and test a set of signal processing and machine learning tools, to support computational diagnosis of dementia. We focus on conversational speech data due to its high availability through cellphones and connected devices (i.e., no custom sensors are needed) and its high expressive power (i.e., much can be inferred about an individual’s state from the content and quality of his/her speech). The primary aim of this article is to show that this speech data can provide valuable insight into the presence or absence of dementia and into the specific kinds of dementia if present. Towards this aim, we set two specific goals:

1. To create an algorithm with leading results for determining whether an individual has dementia or not based on recorded speech.
2. To determine the key features needed for this classification.

The methods used in this article follow a three-stage process. First, features are extracted using readily available open-source tools including the openSMILE package for acoustic feature extraction and the Google speech recognition API for text-based feature extraction [18, 81]. Second, feature selection techniques are applied to remove noisy features. This step was originally applied in post-processing to select those features which were most indicative to the clinical team. We later found this feature selection process significantly improved the final classification results and so included it as the second stage of the process. The final stage performs classification by which we undertake both the bimodal task of determining whether an individual is healthy or has dementia and the multimodal task of determining what diagnosis if any an individual should receive.

The system and study presented here was produced with two potential future applications in mind. First, we aim to pave the way toward an early detection mechanism for dementias such as the ones described here. Although the results we present are on a dataset that demonstrates considerable selection bias (i.e., the proportions of dementia subtypes differ from the true population prevalence), results approaching human performance suggest that the proposed techniques could one day be applied to early detection through an easily accessible medium such as a smartphone application. Second, we aim to support general practitioners that may not have specialists nearby to which he/she could refer difficult cases. For instance, we believe that this proof-of-concept demonstrates the potential to facilitate distinction between diseases that typically required special training to distinguish (e.g., bvFTD vs. PPA).

Outline

The rest of the article is organized as follows. Section 2 gives background on speech processing and dementia and details related work in automatic dementia detection. Section 3 discusses the feature extraction process, describing the dataset and the collection of acoustic and text-based features. Section 3 discusses the classification process including methods for feature selection and classification. Section 4 describes the results obtained. Section 5 discusses the results and limitations of the work. Section 6 provides some conclusions on the work and possible future directions.

3.3 Background and Related Work

Types of Dementia Relevant to this Work

Dementia is defined clinically as a progressive cognitive disorder that leads to an inability for an individual to independently perform their activities of daily living. While many view dementia as synonymous with Alzheimers disease, there are in fact several forms of dementia. *Alzheimers disease* is most common over the age of 65, but dementia can strike younger people, often resulting in misdiagnosis. In people under the age of 65, dementias

can be mistaken for personality changes or a psychiatric illness such as depression [78]. Even if a dementia is suspected, the wider variety of possible dementia types in younger age make a precise diagnosis difficult.

Frontotemporal dementia (FTD) at least as common as Alzheimers disease in people under the age of 65. There are three main forms of frontotemporal dementia: *behavioral variant frontotemporal dementia* (bvFTD), *semantic variant primary progressive aphasia* (svPPA), and the *nonfluent variant of primary progressive aphasia* (nfvPPA). All FTDs interfere with social interaction: bvFTD causes a loss of social and emotional regulation and appropriate interaction, whereas the two forms of primary progressive aphasia interfere with language comprehension and production [58, 57].

Accurate diagnosis of dementia, including the subtype, can have important implications for treatment and prognosis of the disease. This will only become more important with the advent of new therapeutic agents, as there is a growing recognition that treatments are most likely to be effective early in the disease course, therefore requiring early diagnoses.

Background on Speech Processing

Methods in computational processing of speech have advanced considerably in recent years. Private companies have developed state-of-the-art *automatic speech recognition* (ASR) schemes by leveraging massive quantities of labeled training data. With these large datasets, deep learning techniques first replaced more classical *Gaussian mixture models* (GMMs) for recognizing individual phonemes then *hidden Markov models* (HMMs) for modeling temporal probability distributions.

In limited data regimes where deep learning methods are prone to overfitting, however, GMM-HMM techniques continue to provide cutting-edge results. These techniques typically represent acoustic data by Mel Frequency Cepstral Coefficients (MFCCs) or linear spectral pairs (LSPs) and their first or second temporal derivatives. The purpose of this preprocessing is to define summary statistics of the raw acoustic waveform that are smaller in size by selecting the information relevant for making accurate discrimination between the sounds to which humans are sensitive. This preprocessing reduces the feature space from having dimension in the hundreds of thousands (e.g., a 10 second window sampled at 44kHz provides 440,000 data points per sequence) to a low dimensional manifold in which we expect the relevant information to occur. In this way, efficient calculations over acoustic data can be performed while minimizing the loss in expressive power.

Related Work

Recent studies in automatic dementia detection have focused on extracting content based features and training simple classification algorithms.

In [70], they use the ACADIE corpus of transcribed conversations of AD patients compiled within a study of donepezil. Using the frequencies of common words in the text, they achieve 95% accuracy in detecting AD. [32, 33] use Carolina Conversation Collection

composed of both raw conversations and transcripts . [33] mixes lexical richness measurements with hand designed features including filler words, repetitions, incomplete words and go-ahead utterances.

Some studies have intended mixing textual features to some acoustic measures to predict specific dementia types detection. [64] adds few acoustic measurements to text based features but focus on detecting trouble-indicating speech for subjects already with AD. [23] focus on PPA detection using audio of patients asked to tell the Cinderella story and its transcribed version made by research assistants. By using frequency text based features as well as pauses and fundamental frequency variations, they achieve 87% accuracy in detecting PPA on a dataset of 40 people. [51] uses a combination of part of speech related features and pauses to discriminate different variants of frontotemporal lobar degeneration on 38 patients.

In our study, we focus on a fully automatic dementia detection procedure. We use a consequently larger dataset with 124 individual after preprocessing and including various subtypes of dementia. Thus, the high accuracy presented here is less likely due to sample bias or overfitting than previous results presented in the literature. Our main objective being early detection, we used patient data for early dementia variants. By using only raw conversation, we designed algorithms easily applied in real world context (phone conversations, recorded appointments). We apply a more statistical approach using a large combination of acoustic based features (frequencies, pitch, loudness, pauses) and textual based features (word frequencies, richness, word similarities, reaction times). By looking at different measurements (accuracy, recall, precision, importance of features), we provide an analysis of how we could generalize with bigger datasets and what kind of measurements could be of interest for practitioners.

3.4 Feature Extraction Process

Description of the Dataset

The dataset was obtained by gathering recordings of couples (participants with dementia and a familial caregiver). Patients were diagnosed with bvFTD, svPPA, nfvPPA and eoAD by a team of neurologists, speech pathologists and neuropsychologists. BvFTD diagnoses were determined using the Neary clinical criteria [48], and svPPA and nfvPPA by consensus criteria [31]. AD was diagnosed using National Institute on Aging-Alzheimer’s Association diagnostic guidelines [41]. The patients were in early stages of dementia as judged by a mean Mini-Mental State Exam score of 23.4 (SD 5.8) [22]. All assessments were conducted between 2002 and 2014. All study participants provided written consent regarding study participation. The study was approved by appropriate institutional review boards. Patients with bvFTD, svPPA, nfvPPA and eoAD (early onset Alzheimer Disease) were evaluated along with a healthy companion, usually a close friend or family member.

The dataset used for this research consisted of 98 audio conversations between an individual with dementia and his/her close caregiver, obtained during an assessment of emotional

functioning conducted at the Berkeley Psychophysiology Laboratory at the University of California, Berkeley. Laboratory procedures for obtaining samples of conversations were derived from those originally developed by [44]. Couples were instructed to discuss a mutually selected topic of continuing disagreement in their relationship. Each conversation lasted between 10 to 15 minutes. Recordings of the conversations were obtained using unidirectional Shure lavalier microphones attached to each participant.

The audio from the conflict conversations was then transformed into .wav files. A spectral noise gating algorithm was used to remove background noise in Audacity 2.0.3 [69]. Trained research assistants blinded to speaker diagnosis labeled controls and participants speech in Praat [12], an acoustic analysis program. Environmental noises and non-speech sounds were labeled for exclusion. Each labeled conversation was checked for quality before use. A Praat script then extracted intervals of uninterrupted speech for each speaker in the conversation. These intervals were then subjected to further analysis.

In addition to the raw audio files, the times each speaker concluded or began were manually marked in a file that we will call textgrid. Marking times in this fashion allowed for simple segmentation of the conversation. To ensure quality, each timing file was independently checked and verified. Demographics for 126 of the individuals contained within the sample population were also provided. Two of the individuals in the sample spoke so little that no analysis was possible. Thus, from the original 98 audio conversations, segmented audio from 124 individuals was extracted and analyzed alongside matching diagnoses and demographic information. The others couldn't be used because the diagnosis was not provided.

The sample characteristics of the dataset are shown in Figure 3.1. The dataset shows minor bias toward *healthy controls* (52.4%). Those controls were obtained by taking the healthy caregiver speech in the conversation. Within the dementia subsample, the set is divided among three classes of disease: *Alzheimer's disease* (AD, 12.7%), *behavioral variant Fronto-Temporal Dementia* (bvFTD, 15.9%), and *Primary Progressive Aphasia* (PPA, 18.3%). The gender distribution is 51% male, 49% female. The age distribution is approximately Gaussian centered at the 60-65 age bracket (Figure 3.2).

While conclusions drawn here may not generalize to the whole population due to the inclusion of a higher proportion of less common subtypes of dementia (e.g. bvFTD, PPA), the diversity of the dataset enables better accuracy for less frequent subtypes as well as better detection in a relatively young population.

Acoustic Features

Vocal production can be influenced by social, emotional, autonomic and motoric processes, all of which may be altered by neurodegenerative illness. This has led to demonstrated differences in vocal production from healthy controls in a wide range of neurological illnesses such as Alzheimer's disease. That vocal production can be measured by several measures. For example, the Mel-frequency cepstral coefficients (MFCCs) – popular for autonomic speech processing tasks – use a scale accounting for human hearing perception. It is obtained by

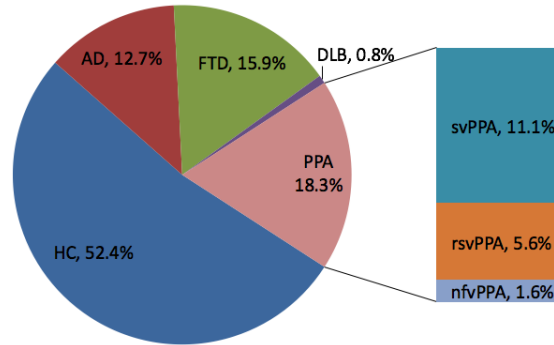


Figure 3.1: Distribution of the 126 individuals with respect to disease. The set comprises 66 *healthy controls* (HC, 52.4%), and 60 individuals with *Alzheimer's disease and related Dementias* (ADRD, 47.6%). Of the affected individuals, the primary diagnosis for 16 is *Alzheimer's disease* (AD, 12.7%), for 20 is *behavioral Fronto-Temporal Dementia* (bvFTD, 15.9%), for 1 is *Dementia with Lewy Bodies* (DLB, 0.8%), for 23 is *Primary Progressive Aphasia* (PPA, 18.3%). Within the PPA segment, 14 show the *semantic variant* (svPPA, 11.1%), 7 show the *right semantic variant* (rsvPPA, 5.6%), and 2 show the *non-fluent variant* (nfvPPA, 1.6%)

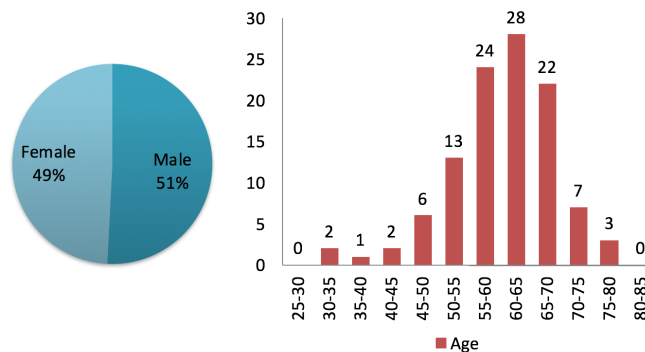


Figure 3.2: Distribution of the 126 individuals with respect to gender and age.

mapping the power spectrum of the sound onto mel scale bands via equidistant triangular overlapping filters, then taking the logarithm of the powers within each mel frequency band. Line spectral pairs (LSPs) are a representation of linear prediction coefficients (LPCs), which themselves represent transmissions of the spectral envelope of speech. Compared to LPCs, LSPs have relative high stability and low sensitivity to quantization noise. [66]

In recent studies of computer-aided dementia detection [32, 33, 23, 25, 51], the study of speech characteristics plays an important role but is generally focused on content. Specific acoustic features stated above have been used partially in some studies [23, 25, 51] and have been proven significant in similar tasks such as emotion recognition [5] or autism detection [40].

To gather a wider variety of measurements, we computed the large acoustic feature set

using the open-source tool, openSMILE [18]. We created, given a part of the audio conversation, 26 indicators for every frame of the audio, including principal frequency, MFCC, LSP, loudness and voicing probabilities. By optionally applying the discrete differentiation (finite difference quotient) followed by 19 aggregating functions such as mean, standard deviations, quantiles or regressions slope and offset, we obtained a set of 988 features.

To compute the aforementioned features, we first divided the pre-filtered audio files into samples of length at least 8 seconds containing speech from only one person (between 4 and 20 samples per person). This ensured that each sample is long enough to be relevant. We computed the features on each sample and then took the mean on all the samples extracted for the same person from the same conversation. We additionally stored 8 features on the whole conversation, including the segments shorter than 8 seconds. We refer to these as conversational features because they encapsulate more global information relating to the tone of the conversation including ratio of speech, mean length of utterances, and the number of uninterrupted parts of speech per conversation.

Following the study from Pakhomov et al. [51] and other similar studies, [23, 25], we finally added 5 features to describe the pauses in the speech (functions of length and number), using a custom pause extraction process that finds pauses by merging close sets of consecutive silent frames. We extract pauses of length more than 1ms, which correspond to what the human ear can detect. The resulting dataset of acoustic features has 1001 elements.

Textual Features

Given the good results of classification on content-based features [32, 33, 23, 25, 51], we also extracted textual features. We started with an automatic transcription with a sufficient word-by-word accuracy (see Appendix C). The text was then modelled as a bag of words, meaning the order of the words was ignored and the text was viewed simply as a set of with multiplicity. The general process is therefore to apply text-based metrics on each word separately and then use four aggregation functions (min, max, mean and standard deviation) to derive features in a similar manner as acoustic features generation. In order to maintain feature relevancy, we provided a specific function for each measurement to test if every word was relevant. For example, in bigram frequencies of the letters, we did not consider words of length less than 3. We then stored the ratio of words not relevant in a feature: the relevance ratio (figure 3.3).

A large number of the word measurements were selected from the Elexicon project [7] which provides behavioral and descriptive data on 40,481 words. The corresponding features have been either computed or collected among six universities on normal students and staff. From this, 39 measurements were used divided in six categories: word complexity (frequency in different corpuses), orthographic neighborhood (size and complexity of words close in spelling), phonographic neighborhood (close in sound), numbers of phonemes/syllables/letters, bigram frequencies (mean frequencies of consecutive pairs of letters) and reaction times to speeded naming and lexical decisions (time to pronounce a word and to identify that a combination of letters is a word). For these features, the relevance

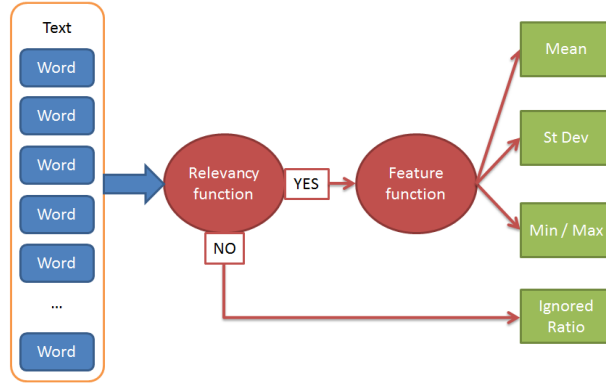


Figure 3.3: Vocabulary Features Extraction Process. Each word considered relevant is used to compute aggregation functions. The others are counted to compute the relevance ratio

function was set to accept all words present in the Elexicon project. Plural and conjugate forms were assigned the value of the root word.

Eight other measurements were used in addition: bigram and trigram frequencies of words not included in the Elexicon project (4 features), vowel and consonant distributions (2 features), and the age of acquisition (2 features) [43]. The age of acquisition has been used in several similar studies [24, 25, 23] and originates from a database that provides the mean and the standard deviation of the user-reported age at which users learned each word. Data for this database was collecting using Amazon Mechanical Turk, the web-based crowdsourcing technology. The relevance function was therefore set to reject unavailable words as well as those with high standard deviation (greater than 4).

In addition to vocabulary oriented text features, inspired by the work of [13, 32, 33], we also added four features that measure the richness of the vocabulary: number of words (N), richness ratio(RR), Brunet Index(BI), and Honore Statistic(HS). They are defined as:

$$\begin{cases} N \\ RR = \frac{V}{N} \\ BI = N^{-0.165V} \\ HS = \frac{100 \log(N)}{1 - V_1/V} \end{cases}$$

where N is the number of words, V the number of distinct words and V_1 the number of words used exactly once. To the richness features, we also added the TF-IDF indices (Term Frequency – Inverse Document Frequency) [65] for words present in at least 25% of the documents and that are in the top 15 features selected with at least one of the feature selection techniques described below. TF-IDF is defined as:

$$tfIdf_{i,j} = tf_{i,j} \cdot \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

where t_i is the i^{th} term, d_j is the j^{th} document, $tf_{i,j}$ the frequency of t_i in d_j and $|D|$ the total number of documents. This process resulted in 10 words. All combined, a total of 249 textual features were defined.

3.5 Classification Process

Preprocessing and Feature Selection

Once the feature set was computed, we performed several preprocessing steps. We separated the data by gender due to variance in acoustic features for men and women. We discarded every sample with missing acoustic features and took the mean over all samples to obtain a single value for each person. We also discarded every person associated with a transcript with less than 20 words in order to keep relevant vocabulary features. Due to limited data, we did not discard subjects for other reasons. If values were missing for age, we took the mean value. If no information on pauses could be extracted, we set a 0 value indicating no pauses were present. We finally eliminated any feature constant for every subject since these features only extend computation time without providing any discriminative power. After the preprocessing steps, 124 subjects remained in the dataset. To this dataset we applied leave-one-out cross validation.

Because of the high number of features coupled to a small number of samples, feature selection was critical to high performance. Because we wanted to keep interpretability as well as some stability in the set of features selected, we used precomputed scoring functions on features to select the k best ones. On every cross-validation set, the scoring function was computed for every feature and then the average was taken across all sets. This had the primary advantage of being easily interpretable and more stable than simply computing the score functions directly. It is also significantly faster than computing it for each set of the cross-validation individually. However, because it uses the target values of all samples, it risks overfitting. To verify that only minor overfitting occurred, we tested the final algorithm on 10 new subjects and observed a drop in accuracy of less than 5%.

We used three different types of scoring functions for the feature selection. First we used 3 variance-based approaches: ANOVA, the Welch t-test and the Chi2. In brief review, ANOVA tests the probability that according to one feature, individuals from two labels are likely to come from the same population. The Welch t-test takes a similar approach but does not assume equal variance in both labels. The Chi2 tests the likelihood of independence between every feature and the label. We next used the coefficients from both *Lasso* and *Ridge regression* as scoring functions. We tuned the sparsifying coefficient of each to select approximately 100 variables. We finally used the feature importance in several tree-based methods: *Decision Tree*, *Random Forest*, and *AdaBoost* with Trees as base learners using either the entropy or the gini impurity as criterion for the cuts. The methods used were extracted directly from Scikit Learn Python library [54] except for the Welch t-test where the SciPy library [39] was used.

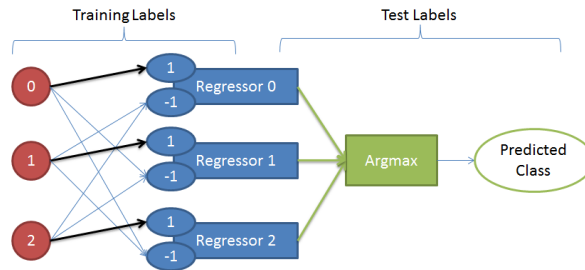


Figure 3.4: Procedure to perform classification using regressors. 0, 1 and 2 represent the labels to predict (HC, AD, FTD, PPA). One regressor is associated to each of them

Classifiers

For the classification itself, we used a large number of standard classifiers from the Scikit Learn Python library [54]. We first used simple linear models such as *Logistic Regression*, *Linear Discriminant Analysis* (LDA), and *Support Vector Machine* (SVM). We also used the Decision Tree classifier and given its good results, we tried different tree-based methods exploring the bias-variance tradeoff, such as *Random Forest*, boosting (*AdaBoost* and *Gradient Boosting*), and *Extra Trees* (a more randomized version of Random Forest). In addition to the above algorithms, based on the good accuracy of neural networks in many speech-based classifications, we implement the *Extreme Learning Machine* (ELM) [35] which is a one hidden layer neural network with weights between the input and the single hidden layer set randomly. It has the advantage of being significantly faster than traditional multilayer networks and less prone to overfitting given the reduced parameter set. We did not apply any deep networks given the relatively small dataset, and the unavailability in the literature of relevant pre-trained deep networks such as those available for vision through ImageNet [16].

In order to leverage penalization properties (sparsity, handling of highly correlated features) on simple models, we used regression algorithms such as Lasso, Ridge regression, Least Angle Regression, and Elastic Net. To do so, we trained a regression algorithm for each label in the output and used the index of the regression with maximum output as a prediction (see figure 3.4). The probability of belonging to each class was computed using the following formula:

$$Prob(x \in C_l) = \frac{R_l(x) - \min(-1, \min_{i \in L}(R_i(x)))}{\sum_{j \in L} R_j(x) - \min(-1, \min_{i \in L}(R_i(x)))}$$

with C_l the class of label l , $R_l(x)$ the score of the regressor associated to the label l for x , and L the set containing all the labels. With the definition above, the regressor output with the lower value had 0 probability if it is under -1, but had a positive probability otherwise.

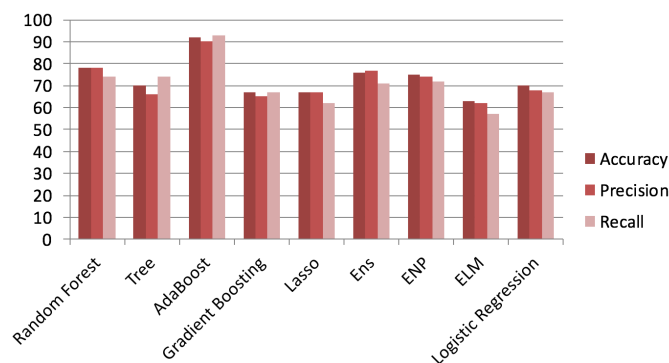


Figure 3.5: The best results in determining whether a speech segment belongs to an individual with dementia or a healthy control. Two-step AdaBoost, or Selective Boosting, demonstrates 92% accuracy and greater than 90% precision and recall.

Finally, we used different Voting Classifiers. As typical in ensemble methods, the idea was to combine the predicted probabilities from different algorithms to get a more stable one which incorporates the benefits in prediction from different algorithms. In order to avoid giving too much importance to some algorithms, the minimal probability was set to 0.01. This threshold also enabled the method to discredit the labels with null probability for one of the classifier while still differentiating if several labels had null probabilities.

3.6 Results

Classification Accuracy

The best classification accuracy scores for the bimodal problem (Dementia vs. Control) are shown in Figure 3.5. The best result of 92% accuracy was achieved using AdaBoost with the 50 best features selected by AdaBoost a priori, as discussed in the previous section. The precision and recall are each over 90%. All other classifiers show significantly weaker results. With this result, the first goal is accomplished of providing leading-edge prediction accuracy. The result also seems to show that a two-step AdaBoost significantly improves accuracy. In the below sections, we refer to this method as Selective Boosting.

The best accuracy scores on the multimodal problem are shown in Figure 3.7. The scores are considerably lower. In the best case, the overall accuracy is 70%, obtained using Gradient Boosting. As shown, this result strongly benefits from the ability to separate healthy controls with greater than 90% recall. Among the three disease types, each presents similar recall suggesting the three variants present similar classification difficulty.

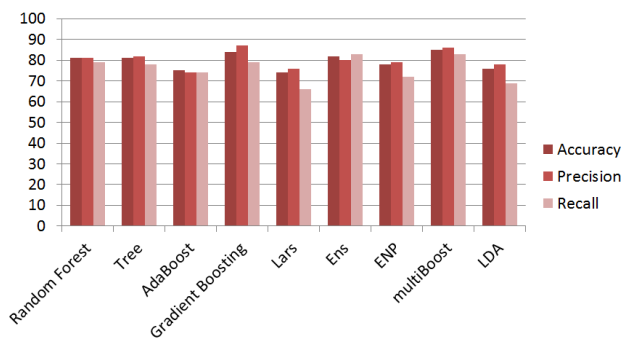


Figure 3.6: The best results in determining whether a speech segment belongs to an individual with dementia or a healthy control. Most tree-based methods reach accuracies higher than 80%.

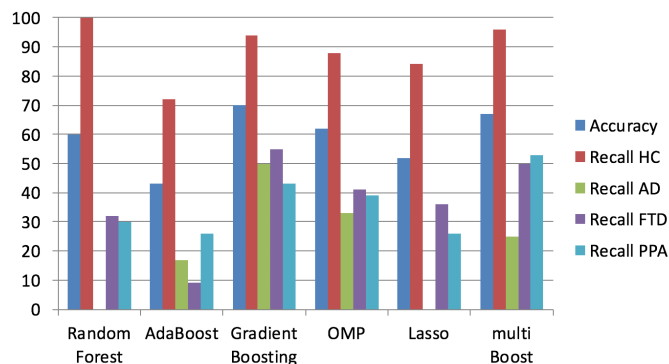


Figure 3.7: The best results in determining the diagnosis if present. Gradient Boosting demonstrates 70% accuracy, greatly benefitted from high recall among healthy controls.

Feature Selection

As shown in Figure 3.8, the impact of feature selection is significant. In this case, feature selection is performed a priori using the AdaBoost scoring function. Thus, the benefits of feature selection are less apparent for the other algorithms. In the case of AdaBoost, the benefits are significant. With 200 features, the accuracy is 74%. The accuracy increases as noisy features are removed and overfitting is reduced until 50 features are used where the accuracy is 92%. After this point, feature selection reduces the accuracy as useful information is lost. When only 5 features are used, the accuracy drops to 76%. Although lower than the peak performance, this result is unexpectedly high given how little information the algorithm has access to with only 5 features present. These features are shown in table 3.8 and discussed in the following section.

To examine the effects of feature selection, we limit each algorithm to only using 15 features. We perform feature selection using a decision tree with Gini criterion to provide

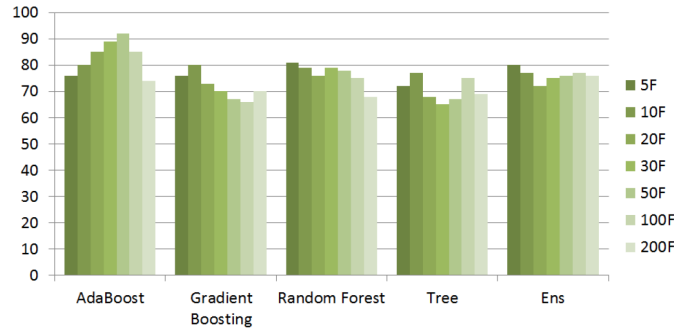


Figure 3.8: The effect of feature selection on the bimodal classification results. Feature selection is performed a priori using the AdaBoost score function, so the change in accuracy for AdaBoost is most indicative.

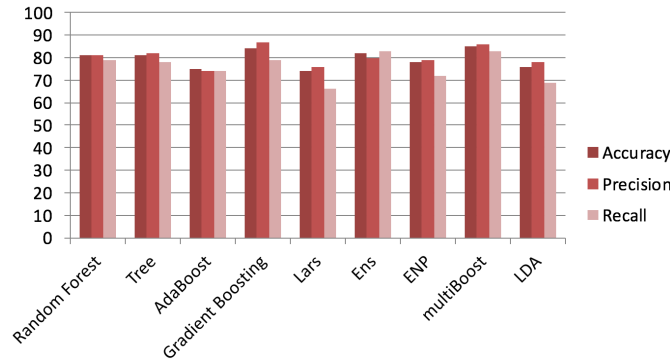


Figure 3.9: The best results when limited to 15 features. Note multiBoost still demonstrates 85% accuracy. Here feature selection is performed using a decision tree with Gini criterion.

similar benefit to each of the tree-based algorithms. The results are shown in Figure 3.9. In this case, the best results are achieved by multiBoost with 85% accuracy. As discussed in 3.2, multiBoost is a Voting Classifier based on the predicted probabilities of different tree-based methods to reduce variance (Random Forest, Extra trees, AdaBoost and Gradient Boosting on Decision Trees). As compared to Figure 3.8, the accuracy of AdaBoost is drastically reduced. This demonstrates the benefit of allowing a specific algorithm to perform feature selection. It also shows that although AdaBoost is able to leverage less informative features to achieve a higher final accuracy, it does not perform as well as the other classifiers with limited information.

Most significant features

The most significant features for the bimodal classification problem are shown in Table 3.1. The most indicative feature is the proportion of the conversation in which the affected individual is speaking. Six of the features are acoustic features based on functions of the

Mel-frequency cepstral coefficients (MFCC) and line spectral pairs (LSP). Interestingly, the remaining features are each based on how difficult a word is to say (e.g, the variance of words used which start with a vowel). Other text-based features such as how difficult a word is to comprehend are not seen.

Table 3.1: Top 10 features for bimodal classification. The symbol \circ denotes composition of functions.

Bimodal Features	
1	Proportion of Conversation Spent Talking
2	MFCC-9 \circ Moving Average \circ Linear Regression Slope \circ Delta
3	Letter Trigram \circ Minimum
4	MFCC-12 \circ Moving Average \circ Linear Regression Y-Intercept
5	First Vowel \circ Variance
6	Orthographic Neighborhood Frequency \circ Mean \circ Minimum
7	MFCC-12 \circ Moving Average \circ Linear Regression Slope \circ Delta
8	LSP-1 \circ Moving Average \circ Slope
9	MFCC-5 \circ Moving Average \circ Interquartile Range 2-3
10	MFCC-9 \circ Moving Average \circ Minimum

Although its performance in the classification is less interesting than tree-based selection, ANOVA selection gives interesting features that have a strong discriminative power for the prediction of dementia. Its lower performance on the accuracy of the classification is certainly due to high correlations in the feature space that can be captured by the tree and cannot be understood with ANOVA. However, the top features are still interesting ones to consider from a medical perspective. The 10 top features have been summarized in Table 3.2: like for tree based selection, it contains three features on MFCC (8th and 9th). It contains two features on voicing probabilities which are generally more powerful in the multimodal problem. We also see trigrams and orthographic neighborhood features in the top 10 features selected by ANOVA. This seems to suggest that words with unfamiliar sounds are not used as frequently by dementia subjects. Finally, average intensity — which is often noted in clinical practice — appears among the top ANOVA features.

The most significant features for the multimodal classification problem are shown in Table 3.3. The most significant feature is based on the zero-crossing rate, a common feature for determining whether or not a sound belongs to human speech. The second feature is based on the orthographic neighborhood. The remaining eight features are all functions of the acoustics. Again, no features directly involving the complexity of a word in comprehension are present.

Table 3.2: Top 10 features for bimodal classification with ANOVA. The symbol \circ denotes composition of functions.

Bimodal ANOVA Features	
1	Letter Trigram \circ Minimum
2	Number of Samples
3	MFCC-9 \circ Moving Average \circ Skewness
4	Orthographic Neighborhood Frequency \circ Mean \circ Minimum
5	Intensity \circ Moving Average \circ Skewness
6	MFCC-8 \circ Moving Average \circ Delta \circ Kurtosis
7	MFCC-9 \circ Moving Average \circ Delta \circ Linear Regression Offset
8	Voicing Probability \circ Moving Average \circ LR Quadratic Error
9	Trigram with Subtl Norm \circ Minimum
10	Voicing Probability \circ Moving Average \circ LR Linear Error

Table 3.3: Top 10 features for multimodal classification. The symbol \circ denotes composition of functions.

Multimodal Features	
1	Zero-Crossing Rate \circ Moving Average \circ Delta \circ Skew
2	Orthographic Neighborhood Frequency \circ Minimum
3	MFCC-5 \circ Moving Average \circ Linear Regression Slope
4	MFCC-3 \circ Moving Average \circ Linear Regression Slope
5	MFCC-2 \circ Moving Average \circ Delta \circ Mean
6	MFCC-8 \circ Moving Average \circ Kurtosis
7	Fundamental Frequency \circ Moving Average \circ Linear Reg. Y-Int.
8	MFCC-5 \circ Moving Average \circ Delta \circ Quartile 2
9	LSP-3 \circ Moving Average \circ Maximum
10	LSP-0 \circ Moving Average \circ Delta \circ Variance

3.7 Discussion

Clinical Relevance

As the population continues to age, interest in accurate diagnosis of dementia is increasing. At this time, clinical diagnosis can demand much time on behalf of caregivers, patients, and

medical personnel, and these demands are expected to grow. A simpler screen that can be performed in the home environment without imposing additional demands on caregivers, patients or medical staff would be a valuable tool. This analysis shows greater than 90% agreement with clinical judgment in the diagnosis of dementia based on conversational speech alone. Given the ready availability of speech samples, it is possible that a similar approach could permit screening for dementia with potential benefits of early detection and following disease progression over time. This sets the stage for future progress in early diagnosis, prognosis with readily available data streams, and inexpensive distinction between similar speech pathologies.

Classification

The high accuracy in determining whether an individual has dementia or not suggests that computer-aided diagnosis of dementia is worth pursuing. Although the results are on a dataset which shows some bias in dementia pathologies, the accuracy is approaching that of a human expert. With significantly more data including diagnoses confirmed by post-mortem histology as well as data from patients across disease types and stages, an algorithm competitive with human experts seems possible. On the task of dementia detection from conversational speech, we present the highest accuracy achieved by an algorithm thus far on a dataset of this size. It should be noted, however, that the dataset is still relatively small for machine learning applications, thus, accuracies may be overestimated. By achieving the first goal of the article to produce state-of-the-art results for determining whether an individual has dementia from recorded speech, we hope to encourage future work in 1) early detection of dementia and 2) support of fine-grained diagnosis by the general practitioner without access to specialists in clinical neurology. The lower accuracy on the multinomial classification problem suggests that continued research is needed before a computational diagnosis can be performed independently.

Feature Selection

The most discriminative features are related to the proportion of the conversation spent talking, the pronunciation complexity, and the *Mel-frequency cepstral coefficients* (MFCC). The potential significance of a decreased proportion of conversational time spent talking has been noted by clinicians. Although features capturing how difficult a word is to say appeared, no features appeared which are used to describe how complex a word is conceptually, such as the age of acquisition. This suggests that in this cohort of early stage patients, mental capacity is diminished less than muscular ability to perform difficult articulation. The presence of many MFCC features, which are often used in speech recognition algorithms, suggests that the quality of the speech can be very indicative of the underlying disease state. The presence of these features also suggests that the algorithm could be improved by incorporating advances that have been made in recent years to improve upon hand-engineered features like the MFCCs by learning features from speech data itself (i.e., deep learning). Although we

chose not to use deep-learning approaches here based on the paucity of data, future attempts could be made to tune a pre-trained deep speech network to the data presented here.

With key features needed to characterize dementia-like speech, the approach presented here can pose a privacy preserving, minimally invasive method for monitoring disease progression. By extracting the features studied and identified by our approach from regular cellphone conversations, clinicians and family may obtain unique insight into the progression of the disease over time. Rather than saving the recorded conversation itself, these features could be extracted locally, providing little insight into the actual content of the conversation in a privacy preserving architecture. These same features could then be used to perform regression whereby an individual may appear more dementia-like following a change in medication, primary caregiver, place of residence, etc.

Limitations of the study

The dataset could be improved by including more participants, including participants from varying stages of disease, providing data from the same participants over time, and obtaining data with true labels based on post-mortem histology rather than clinical diagnosis. In particular, the dataset analyzed here contained only individuals presenting early dementia pathologies. In order to generalize these results to general diagnosis, the methods should be applied to a broader dataset. It will likely not be possible to claim an algorithm is capable of surpassing the performance of a human expert until a dataset is created which contains labels based on post-mortem histology. It will not be possible to claim an algorithm is capable of surpassing the performance of a human expert until this time. A more interesting goal, however, may be to see how an algorithm can be developed to support the clinical expert to improve the final diagnostic capability as discussed next.

The methods used could be improved by limiting overfitting and improving transcription. To that extent, one could perform the feature selection directly inside the cross-validation. It would complicate the computation of top features and increase the computational cost but would limit the likelihood of overfitting. By improving transcription, one could have better accuracies with textual features, and add measurements on the sentences structure. Moreover, if a team of clinical experts created a list of the features used in diagnosis, an individual clinician could provide a score for each of these features on making a new diagnosis. These feature scores could be used in tandem with the features selected here to perform final classification through standard machine learning techniques. This strategy would avoid the difficult technical hurdle of automatically detecting nuanced features such as changes in eye movements while allowing for the inclusion of features which are difficult for a human to measure such as the frequency with which certain word types are used (e.g., orthographic neighborhood). Existing methods could be further improved by more extensive hyper-parameter optimization (i.e., through random search) and by including pruning techniques into the tree-based learning algorithms which are not included off-the-shelf from Scikit Learn Python library.

Given the size of the dataset available for the present work, several design decisions were made that could be improved upon with the addition of more data. The results are provided based on leave-one-out cross validation. A better indicator of generalizability would be to reserve a strict holdout set containing 20% of the data. This cross-validation method is prone to overfitting, but has received recent justification through iterative data analysis techniques such as the thresholdout method.

3.8 Conclusion

This work presents a method for distinguishing the speech of healthy individuals from people with dementia. The method is based on assembling a large vector of features to characterize the speech, then selecting those features that demonstrate the most discriminative power. From this feature selection, we highlight certain features that are more widely clinically recognized, such as the proportion of the conversation spent speaking, and others which are not currently used by clinicians such as the mel-frequency cepstral coefficients. We show that this method provides discriminative power approaching that of clinicians in binomial classification, but only moderate ability to discriminate between Alzheimer's disease, behavioral fronto-temporal dementia, and primary progressive aphasia.

In the future, we believe that this work could be improved in several ways. First, the method could be fully automated by implementing computational segmentation of the speakers in the conversation. This is easily obtainable from cell phone conversations and obtainable with high accuracy in more natural settings if multiple microphones are used and readily available algorithms such as *independent component analysis* (ICA) are applied. Second, the study would benefit from more data, data from each individual from multiple points in time, and particularly from data in which the post-mortem histology is known. Although we show 70% accuracy in predicting the diagnosis, a more interesting result would be the accuracy in predicting the true disease. The predictive features highlighted could be added to clinical procedures in early dementia detection. However, although a physician may have an intuitive understanding that the MFCC provide discrimination based on pitch in a scale approximating the human auditory system, there is not yet a practical clinical system for detecting and determining correlation between certain abnormalities in MFCCs and certain disease types. Thus, the physician cannot yet use these features to inform their own diagnosis in the same way they can use cues obtained naturally by a human expert in practice such as hand tremor or eye movement.

We hope that as medicine shifts from a reactive to a proactive paradigm, the present work demonstrates a proof-of-concept process that recorded speech provides a data source that is both easily obtainable and presents high expressive power. Moreover, by highlighting the most influential features of the data, we propose a privacy protecting method for performing daily prognosis and suggest methods in which features that are best detected by humans and features that are best detected by machines can be used together to improve the overall quality of care.

Chapter 4

Fall Detection through Video Analysis

4.1 Chapter Abstract

We study robust fall detection in the context of images collected in the light with standard RGB sensing and images collected in the dark with IR sensing. We collect a data set in which 4 healthy adults simulate falls in the home environment. The data set contains 103,315 images in the light and 43,485 images in the dark with 30,608 fall images in the light domain and 10,842 fall images in the dark domain. We explore three methods for domain adaption none of which have previously been explored in the context of fall detection: (1) tuning the pre-trained VGG network to the fall-detection task [68], (2) applying the domain confusion loss developed by Tzeng, et al. [74], and (3) implementing a novel domain-specific data augmentation technique based on the deep style work of Gatys, et al [26].

The best results for our application indicate 0.92 precision and 0.86 recall in the light domain and 0.72 precision and 0.63 recall in the dark domain, both originating from simply tuning VGG. For future work, we will generate a larger data set in a 3-month pilot study, extend the discrete domain adaptation results to continuous domain adaptation under day-light cycles, and explore the use of recurrent neural networks to exploit the time-dependencies between video frames for better fall detection.

4.2 Introduction

Fall accidents account for 26% of all Alzheimer’s related hospitalizations [2] and are thus a major concern and key cost contributor. Unfortunately, safety products developed for falls require a wearable device; they were developed for cognitively aware adults and not designed specifically for individuals with Alzheimer’s disease and related dementias. No dementia-friendly fall detection solution currently exists to affordably provide home AD care within a comprehensive framework.

Our team proposes a system which uses off-the-shelf wall-mounted cameras and wireless sensors to passively detect the key safety concerns for individuals with Alzheimer’s disease.

The proposed system does not require any action of individuals or their caregivers such as wearing a fall pendant and is therefore well-suited for individuals with dementia. Although the proposed system provides several functions critical to AD care, we focus here on the critical issue of fall detection from video.

Prior work in vision-based fall detection [19, 36, 46, 80]

follows generally the same process. The interested group collects a small data set of falls which is necessarily limited given the rarity of fall events and the difficulty for an actor to replicate an authentic fall event. The group then proposes a method which is based generally on a three-stage pipeline:

1. Detection of the person within the frame
2. Extraction of key features from the detected region
3. Classification into fall / no-fall based on these features

In the literature, Stage 1 is often accomplished by simple background subtraction under the assumption that the only movement in the frame is due to human motion. Stage 2 is accomplished through numerous techniques including Gabor feature extraction, ellipse fitting to the human profile, projection histograms, Gram-Schmidt orthogonalization, nonlinear PCA, collections of heuristics, and deep features extracted from neural networks. Stage 3 is performed by applying traditional SVMs, hierarchical SVMs, shallow neural networks, and deep neural networks [19, 36, 46, 80].

Although there is clear room for improvement in Stage 1 using algorithms which can be tuned to person-specific localization such as fast-RCNN [28], the focus of this work is on Stages 2 and 3. Namely, we provide the first application of pre-trained deep networks to fall classification using VGG [68] pretrained on ImageNet [16], we study the robustness of this classification to a change in the image capture modality from day-time RGB sensing (*light*) to night-time IR sensing (*dark*), and we explore two supervised learning techniques for maintaining robustness across modalities. The first method applies the deep domain transfer techniques developed by Tzeng et al. [74], and the second applies the deep style techniques developed by Gatys et al. [26] to learn a domain transfer mapping for domain-specific data augmentation.

4.3 Methods

In this section, we first describe how we generated a fall dataset. Then, we outline the network architectures and training methods used to perform fall detection in multiple domains. All experiments were implemented using Caffe and deployed on a NVIDIA Titan X GPU [37].

Data Collection

In order to train and evaluate our model, we built a small fall data set. We recorded roughly 1 hour of video containing four individuals in a standard living room environment simulating

typical fall behavior. To create a realistic domain shift, we recorded data in a day-time setting with a color camera and in a night-time setting with infrared (IR) cameras. We created bounding boxes and fall labels for human figures with Amazon’s Mechanical Turk, using the Video Annotation Tool from Irvine California (VATIC) [75]. Workers were given instructions on how to select regions they believed contained humans and on how to label what they thought of as a fall or someone on the ground. Images were labeled as “fall” or “no fall”, with a fall defined as a person lying on the ground and not in an intermediate pose. To control the quality of the region proposals and labels, we fine-tuned results from the workers, discarding errant proposals and trimming bounding boxes.

We generated 103,315 light and 43,485 dark data-points this way with 30,608 falls in the light domain and 10,842 falls in the dark domain. We show examples of data from the two different domains in Figure 4.1.



(a) Example of “fall” in the night-time setting.

(b) Example of “no fall” in the day-time setting.

Figure 4.1: Examples of data from the day-time and night-time settings.

Baselines

Our first attempt at performing fall detection was to simply finetune two distinct VGG-16 nets [68] for the light and dark domains. The baseline nets were initialized from the pre-trained VGG-16 nets; only the final layer, fc8, was not initialized with weights from VGG. For training, we locked the weights for the first 6 layers and only allowed the two final fully connected layers (fc7 and fc8) to update weights. The Stochastic Gradient Descent (SGD) solver parameters used for all experiments in this paper are listed in Table 4.1. For the baselines, we noted convergence by 20000 iterations.

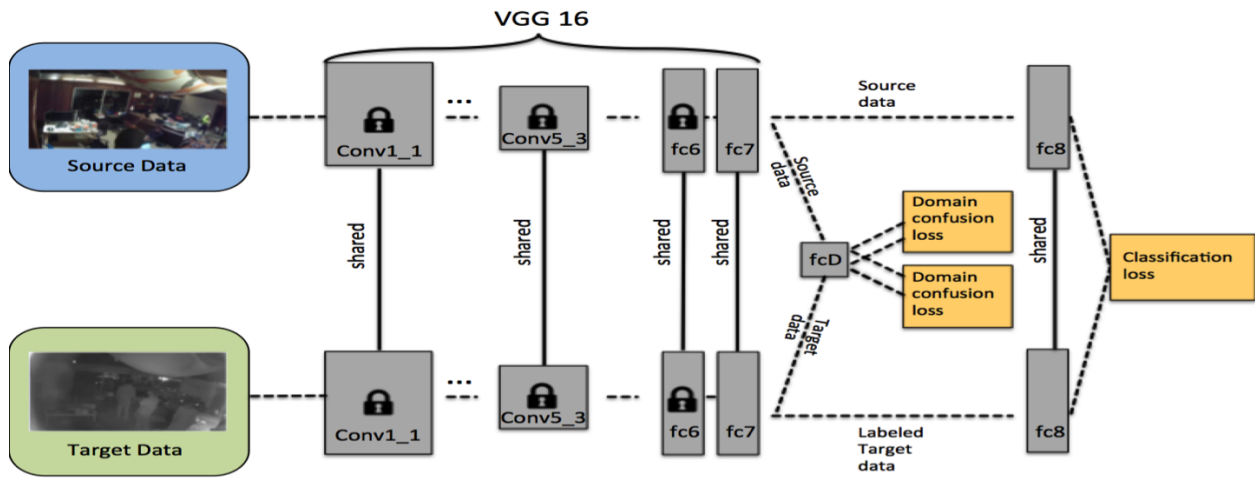


Figure 4.2: Domain confusion net, based on [74], used for experiments. Note that the first seven layers are initialized from the VGG weights [68]. We lock the weights for all layers except fc7 and fc8. In implementation, we use two fcD layers with shared weights to connect to light and dark fc7 layers, respectively.

Batch Size	64
Base Learning Rate	0.01
LR Policy	Step
Step Size	5000
Momentum	0.9
Weight Decay	0.0005

Table 4.1: SGD solver parameters used to train all nets.

Domain Confusion

After performing the baseline experiments, we wanted to see if we could discover a domain-invariant representation to allow use of a single net to perform fall detection in both light and dark domains. We hoped this would allow the accuracy in both domains to be improved by leveraging all available information. In [74], the authors achieved domain transfer through maximizing domain confusion and transferring task correlations from a source domain to a target domain. This method results in a feature representation that is difficult to classify by domain but simple to classify by category, with categories that were close to each other in the source domain representation still close in the resultant domain-invariant feature representation.

We used this technique with a few modifications. Since our problem is a binary classification task (i.e. fall detection), we did not apply the soft label loss to achieve task transfer; we only used the domain confusion loss. In addition, we used VGG as the base net for feature

Model	Precision	Recall	Number of “Falls” in Test Set	Number of “No Falls” in Test Set
Light Baseline	0.859	0.920	2506	18157
Light Domain Confusion	0.884	0.840	2506	18157
Dark Baseline	0.715	0.632	1542	7155
Dark Domain Confusion	0.939	0.511	1542	7155
Style Transfer	0.558	0.640	1542	7155

Table 4.2: Fall detection results for baseline, domain confusion, and style transfer methods.

Model	Precision	Recall	Number of Light Images in Test Set	Number of Dark Images in Test Set
Light Domain Confusion	0.424	0.463	20663	8697
Dark Domain Confusion	0.457	0.409	20663	8697

Table 4.3: Dark domain detection results for domain confusion method. The dark domain detection results in this table correspond to the snapshots used to evaluate fall detection in Table 4.2. Note that the test set used for dark domain detection is the same test set used in Table 4.2 but is partitioned by domain rather than by category.

representation rather than AlexNet [42].

A graphical representation of the domain confusion net is shown in Figure 4.2. Note that we feed in both the light and dark data simultaneously; each input layer has a batch size of 64, respectively. In addition, the DomainConfusionInnerProduct layer (i.e. fcD in Figure 4.2) was provided by the authors of [74]. It implements an iterative update for back-propagation, which is explained in the original paper. For this layer, we used the recommended loss weight of 0.1 for both domain classifier and domain confusion losses. We noticed convergence after 30,000 iterations.

Style Transfer

Gatys, et al. demonstrate that the content from one image and the style from another can be merged by extracting deep features from each image and matching first-order statistics from the content image with second-order statistics from the style image [26]. One example is shown in Figure 4.3. The result is achieved by starting with a white-noise image and

iteratively minimizing the Euclidean loss between the exact features of one layer from the content image and the Gram matrix of the features from several layers of the style image. For example, Figure 4.3 is achieved by extracting the features from one feed-forward pass of the original VGG network pre-trained on ImageNet, then matching the content reconstruction from convolution layer 4_2 with the style reconstructions from convolution layers 1_1, 2_1, 3_1, 4_1 and 5_1.

In this work, we apply the deep style algorithm to domain-specific data augmentation. Given that training data is present in both domains, we propose to learn a mapping from one domain to the other whereby data from both domains can be leveraged to improve overall accuracy. Given that the dark domain by definition captures less information, we propose to map all data from the light domain into the style of night-vision capture. Although the opposite direction is also possible, we show in Figure 4.4d that attempting the opposite direction is an ill-posed problem due to the relative lack of information in the dark domain.

Style transfer was achieved by applying the Gatys style transfer algorithm on a frame-by-frame basis based on code developed by [49]. To ensure diversity in the data augmentation scheme, frames from the light domain were randomly matched with frames from the dark domain. The intent of this scheme was to develop an augmented data set matching the global statistics of the night vision data set rather than the specific statistics of a single frame. This algorithm was parallelized and implemented using an NVIDIA Titan X GPU where 8 transformations could occur simultaneously. Each individual image transform was limited to 200 iterations. With this implementation, the processing time required to transform 30,000 images was 8 days. Due to the limited time remaining for the deadline, the transformation was halted after 7,500 transformations from which 27,645 cropped images were extracted. This augmented data was added to the existing training set containing 34,788 images from night-vision capture for a total augmented data set containing 62,433 images.

4.4 Results and Discussion

In this section we discuss the results from our domain confusion and style transfer experiments. The key results are displayed in Table 4.2. We measured the precision and recall of the different approaches with a *positive* result corresponding to a fall. The results reveal that domain confusion produces a classifier that has fewer false alarms (i.e. better precision) at the expense of more missed detections (i.e. lower recall). Alternately, our results from style transfer show minor improvements in recall at the expense of dramatically reduced precision.

Baseline vs. Domain Confusion

There is an interesting trade-off between the baseline model and the domain confusion model. Recall was higher for the baseline model for both the dark and light domains by roughly 10%. In contrast, precision was higher for the domain confusion model for both domains. This suggests that the models that are trained only on a single domain are more sensitive fall



Figure 4.3: An example of deep artistic style transfer from [26] whereby the content of image A is transformed into the style of 3 separate paintings in images B, C, and D.

detectors. However, the models that are trained on both domains concurrently and exhibit domain confusion are better at rejecting “no fall” cases than the single domain models.

In this application, we are more sensitive to not missing falls. However, too many false alarms could overload supervisors tasked with deploying resources in an emergency situation. In the light domain, the baseline model is preferable as there is higher recall with relatively low loss in precision. However, in the dark domain, there is a sizeable drop in precision if we choose the dark baseline model. Thus, we do derive value from domain confusion for detecting falls in dark environments and further work is warranted.

To verify that the domain confusion worked as expected, we looked at the domain confusion net’s ability to distinguish between light and dark domains at different training iterations. Table 4.3 shows the precision and recall of the domain confusion net at the iteration matching the best performance in light and dark domains (i.e. the same set of weights used for the results in Table 4.2). In this context, we designate the dark domain as a *positive* result and the light domain as a *negative* result in calculating precision and recall. Figure 4.5 shows that the nets quickly learn how to tell the domains apart but then converge to a result where both precision and recall are under 50%. Therefore, we conclude that the domain



(a) Original picture from the dark domain.



(b) Transformed picture with content of light domain (4.4c) and style of dark domain (4.4a).



(c) Original picture from the light domain.



(d) Transformed picture with content of dark domain (4.4a) and style of light domain (4.4c).

Figure 4.4: Examples of style transfer with originals on the left and transformed images on the right.

confusion loss is working and that our trained domain classifier is poor at distinguishing the feature representation of a dark image from a light image.

Style Transfer

Figure 4.4 shows the qualitative results of style transfer. When transforming the content from a light image (4.4c) to appear in the style of the dark domain (4.4a), the results appear as expected (4.4b). For example, areas of local brightness are transformed to resemble lights in the dark domain. This can be seen in the bottom left corner of Figure 4.4c where the white spot of table surrounded by dark headphones is transformed to resemble the light emitting from the bicycle reflectors in Figure 4.4b. Similarly, the cloudiness from Figure 4.4a appears in Figure 4.4b although minor separations across color channels occur. When attempting the transformation in the reverse direction (Figures 4.4a, 4.4c, 4.4d), the inability to compensate for information loss becomes readily apparent. For example, in comparing the top right corner of Figures 4.4a and 4.4c, the colors on the canvas are lost by the IR camera.

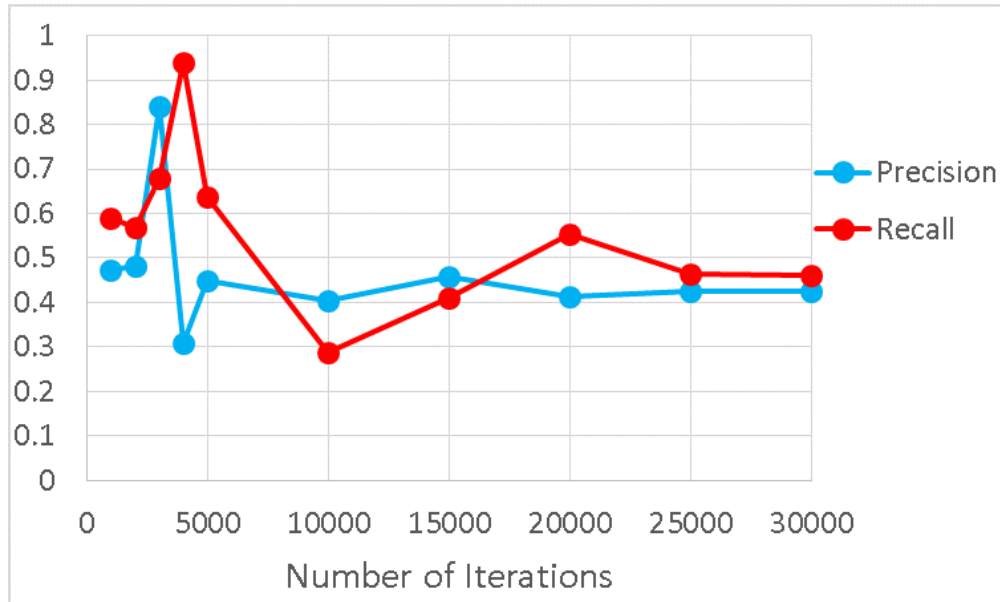


Figure 4.5: Measurement of the domain classifier’s ability to distinguish between light and dark domains over training process. There are initial spikes in precision and recall, followed by convergence to under 50%. Note that the light and dark domain confusion net results in Tables 4.2 and 4.3 occur at 25,000 and 15,000 iterations, respectively.

In the same region of Figure 4.4d, the style transfer mechanism generates a seemingly random accumulation of shapes and colors where it is unable to compensate for the information loss. Interestingly, it is still able to learn selected transformations such as the appropriate color for skin.

As seen in Table 4.2, augmenting the training set in the dark domain with transformed images from the light domain did not provide significant benefit. The recall is improved by 0.008 at the expense of a 0.157 reduction in precision. Although the priority in this work is on detecting all events at the expense of possible false alarms, this great of a trade-off is not warranted. Given that only one fourth of the available light data set was augmented due to time constraints, it remains to be seen how this result scales with the amount of transformed data. It may be that the results continue to improve by continuing to add augmented data, but it may also be that once the majority of the training set originates from simulated data, the statistics from the training set deviate too much from those of the test set to provide satisfactory performance. Similarly, it may be that if standard data augmentation techniques were applied to generate an equal amount of supplemental synthetic images, the results from standard data augmentations may actually perform better than the domain-specific data augmentation performed here.

Successes and Failures

To better understand the representation of the domain confusion net utilized, we took illuminating sample failure and success cases in light and dark domains and used them to identify where improvements could be made.

Figure 4.6 contains sample light images that are particularly instructive. In the left column, we see that Figure 4.6a and Figure 4.6c appear similar, yet our classifier predicts that Figure 4.6a contains a fall while Figure 4.6c contains no fall. To make sense of this, we recall that our net was pre-trained on ImageNet with many human images in upright positions with visible heads. A possible cause of this failure mode may have been that the image in Figure 4.6c contains a barely visible head in a non-standard position, making it difficult to determine if the object on the floor is human.

In the right column, we have two images, Figure 4.6b and Figure 4.6d with fallen humans in similar positions. However, the large occlusion in Figure 4.6d fools the net into making the wrong prediction.

Samples of the dark images provide even more information on success and failure modes of our classifier. In the left column (4.7a and 4.7c), we observe two images that appear nearly identical. Yet, the net correctly classifies one image and not the other. We explored this perplexing behavior further. We found that the difference between both images is that Figure 4.7c has a uniformly lower pixel intensity value. This misclassification could suggest that the image lies directly on the decision boundary, but more likely is a sign that the net is over-fitting to features specific to the training set. The images in the right column (Figures 4.7b and 4.7d) indicate that strong motion blur might also be a reason for misclassification.

4.5 Conclusion

The hope of this study was to develop a method which could be used to implement fall detection with high accuracy regardless of the domain of origin. We extend the prior work in the field by employing a pre-trained deep network for feature extraction, performing classification by implementing the current state-of-the-art for domain adaptation, and proposing one novel method for domain-specific data augmentation. Unfortunately, the results do not yet provide the high performance guarantees needed to provide this system to families caring for a loved one with Alzheimer’s disease as evinced by the low accuracy in the dark domain. In the light domain, however, results are on par with existing wearable techniques for fall detection which show 90% accuracy on realistic datasets [6].

The best results from the light domain indicate 8% of falls will be missed and 2% of non-falls will generate false alarms. Given that the camera sampling rate is 7 frames per second, this corresponds to an average false alarm rate of 8.4 false alarms per minute. Although the distribution of false alarms is not likely to be uniform and false alarms will likely cluster around areas of uncertainty, this false alarm rate remains too high to perform independent fall detection (i.e., without the support of a human observer). More significantly, 8% of frames



(a) Example of a fall labeled correctly.



(b) Example of a fall labeled correctly.



(c) Example of a fall labeled incorrectly. This image is similar to 4.6a except that the human head is not well distinguished. This is a possible reason for failure, as it would be hard to detect the human.



(d) Example of a fall labeled incorrectly. This image is similar to 4.6b except for the occlusion of the subject of interest. This is a possible reason for failure.

Figure 4.6: Success and failure examples in fall detection in the light domain.

containing falls are missed in this test set preventing the current system from providing the high safety guarantees needed for a home safety system.

In the dark domain, neither domain confusion nor domain-specific data augmentation provide significant improvements over the baseline detection results. Moreover, the baseline results leave much to be desired where 36% of falls are missed and 11% of non-falls generate false alarms.

To improve on the current work and to better understand the problem, we will be generating a larger data set through a 3-month pilot study where falls will be observed under natural conditions. In this study, we will further investigate how these results generalize to continuous domain adaptation given daylight cycles present in normal home conditions, and we will explore oversampling to adjust for class imbalance. We will also explore the use of recurrent neural networks to leverage the time-dependencies inherent in fall-detection to provide a more natural, accurate, and robust classifier. Finally, we will explore how domain-specific data augmentation changes with more data and how it compares to standard data augmentation techniques.



(a) Example of a fall labeled correctly.



(b) Example of fall labeled correctly.



(c) Example of a fall labeled incorrectly. This image is similar to 4.7a except for uniformly lower intensity values. Incorrect classification could be a sign of over-fitting



(d) Example of a fall labeled incorrectly. Strong motion blur obfuscates the image, making detection harder. Figure 4.7b, a similar image with less blurring, is labeled correctly.

Figure 4.7: Success and failure examples in fall detection in the dark domain.

Chapter 5

Fall Reduction through Video Review

5.1 Chapter Abstract

A camera system composed of off-the-shelf video recording equipment was placed in one 40-resident memory care community for 3 months to collect real-world data regarding the nature in which fall incidents occur. The purpose was to use this data to develop computational algorithms for fall detection. We describe here an unexpected result in which the memory care facility used the fall video to review how incidents were occurring, updated policies and room layout to reduce potential fall risks, and reduced fall rate from 10.5 falls per month on average to 2 falls in the final month. Preliminary analysis shows a statistically significant fall reduction with $p=0.030$. Given the small sample size, further studies are needed and underway to validate this result.

5.2 Introduction

Significance

Fall accidents are the primary cause of AD-related hospitalization, contributing to 26% of all hospitalizations at an estimated annual Medicare cost of \$5.3B [15]. In nursing facilities, individuals with dementia fall 4.05 times/year on average versus 2.33 times/year for other residents [17]. Unfortunately, safety products such as wearable pendants were developed for cognitively aware adults and fail to meet the needs of individuals with dementia that cannot reliably wear or use such devices. Detecting falls early and in an ongoing manner provides significant potential for reduced hospitalization and system-wide savings. Less than 10% of falls lead to serious injury [15], [17], but 50-75% of elderly fallers experience repeat falls [8], [11], [50], [59], [72], [71]. Thus, detecting the first fall and taking preventative action provides significant potential for reducing fall risk. Through a randomized clinical trial of 160 ambulatory fallers, [63] showed that a nurse practitioner analysing a patient and fall circumstances after the event led to 26% fewer hospitalizations and 52% reduction

in hospital days. Rapid fall detection also limits the amount of time fallers spend on the ground. Mary Tinetti, developer of the well-known Tinetti score for fall risk assessment [73], noted the risks of time spent on the ground following a fall in a study of 596 non-injurious falls [71]. Of 313 fallers, 47% were unable to get up independently after at least 1 fall. Fallers who were unable to get up were more likely to die, to be hospitalized, and to suffer lasting decline in activities of daily living (35% vs 26%). These correlations are confirmed in [21], [77].

Related Work

Current commercial solutions for fall detection fail to address how a fall occurred. The most well-known commercial solutions include wearable systems like Phillips Lifeline which demonstrate limited success in dementia care where individuals forget or refuse to wear a device. Non-wearable fall detection systems based on radar and optical sensors are under development by groups including Emerald and C2S, but are not yet commercially available in the US and have not yet demonstrated robustness through evidence-based studies. Fall mats and bed alarms are prevalent solutions in memory care but are intended only for those residents that should never be walking independently. When applied to general fall detection, these alarms suffer from high false alarm rates due to the prevalence of night-time wandering in dementia care. None of these solutions allow care providers to see how falls occur. There is an absence in the academic literature examining how fall review can impact the quality of care. The most relevant study is conducted by Robinovitch et al. [62] in which video is collected from cameras in two long-term care facilities over a period of 3 years and capturing 227 falls. This dataset was collected to determine the most common causes of falls in managed care and so was collected in coordination with care facilities at which cameras were already installed. Video was not reviewed with facility staff with the specific intention of identifying and removing any possible cause. The study thus offers little insight into the effect of introducing cameras or how video review can impact fall rate. It does confirm an increased fall incidence among residents with Alzheimers disease and highlights that 43% of falls captured involved a cause which could be addressed by facility staff including trips, stumbles, hits, bumps, and loss of support from external objects. Many causes such as incorrect transfer of body weight, responsible for 41% of falls, do not provide obvious changes to the environment which staff could address. Continued data collection from this group appears to be in progress [79]. Only one other study conducting video review of falls appears to exist [34] but has significantly smaller sample size and is also based on pre-existing cameras with no feedback to staff.

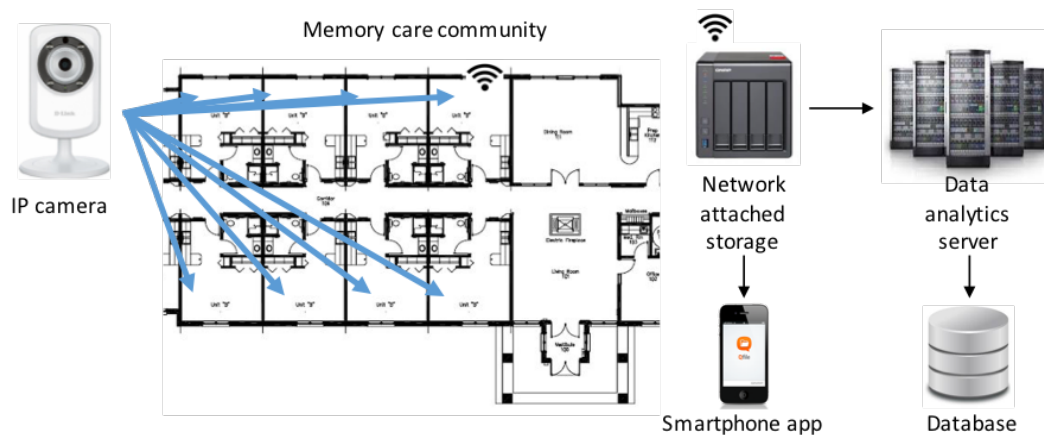


Figure 5.1: Equipment. IP cameras were placed in all common areas and approved private rooms. Video was transmitted from the cameras to the network attached storage (NAS) via Wi-Fi where it was maintained locally for 72 hours after which it was transmitted to a remote server for archiving. Live video and video from the previous 72 hours were made available to facility management via smartphone applications.

5.3 Materials and Methods

Equipment

Figure 1 shows the off-the-shelf video recording equipment used. Cameras were placed in all common areas and private rooms of consenting residents and families in accordance with the privacy guidelines described below. Camera video was transmitted using Wi-Fi to local network attached storage (NAS) devices. Facility Wi-Fi coverage was upgraded using off-the-shelf routers and range extenders to remove Wi-Fi dead zones. Video was maintained on the local NAS for 72 hours before transmitting to a university server where the complete video data set was maintained. A smartphone application was provided for viewing video from the previous 72 hours, developed by the makers of the NAS. A smartphone application for accessing the live video from each camera was provided, developed by the makers of the cameras. Cameras were configured to record only on motion to filter unneeded video and software was developed to support video transcoding and uploading from the NAS to work around bandwidth limitations defined by the upload speed granted to the memory care facility through their internet service provider. The specific equipment provided to the facility included the following:

1. DLink 932L IP camera (x43),
2. QNAP 451+ network attached storage (x2),
3. Netgear AC5300 Nighthawk X8 Wifi Router (x2),

4. Netgear Nighthawk AC1900 Wifi Range Extender (x2).

Fall Review

In the first two months of the three-month study, no video review took place. The original purpose of the study was only to collect video of falls for development of fall detection algorithms. Thus, although video recordings from the previous 72 hours were available to facility management, no formal review occurred. Facility management reported hardly ever using the video feeds during this time due to the many other challenges faced with operating a memory care facility and the little obvious value of the video. After two months, a particularly severe fall incident was recorded. In accordance with procedures approved by the university institutional review board, this incident was reported to facility management. After reviewing this fall, facility management requested reviewing other significant falls. Video was provided and facility management chose interventions which they believed would address the causes. Interventions included movement of furniture that had caused tripping hazards and head injuries and changes to policy that included checking on high-risk residents every hour instead of every two hours at night. The facility did not use any devices for fall detection before or after the incident occurred.

Privacy and Consent Procedures

Privacy and consent procedures were developed with support from the university institutional review board. Permission was granted from facility management to place cameras in common areas and to speak at town hall meeting to introduce the idea to families. After speaking, interested families volunteered their contact information for follow-up discussion regarding cameras within the private room. At the discussion, the study was explained in plain English, and for those families who decided to participate, surrogate consent for the resident living in the care facility was obtained following university guidelines for surrogate decision makers. Residents were required to give assent; if they ever expressed a desire not to take part in the study or have cameras placed in their rooms through verbal or non-verbal communication, they were not included. Video segments defined as improper by the review board were deleted including any video containing sexual activity, actions that could imply abuse if taken out of context, and other incriminating behaviours. Before deleting, the dementia care nurse on the team was responsible for determining if the matter should be taken to facility management or to adult protective services. Following California state guidelines, audio recording was disabled and signs were posted visibly on the door of each private room in which video recording occurred. Before publishing video in any way, media release forms were signed from individuals contained in the videos or from their surrogate decision makers allowing for public release of the specific videos in question. The number of falls recorded and resident population for each month were determined by interview with facility management.

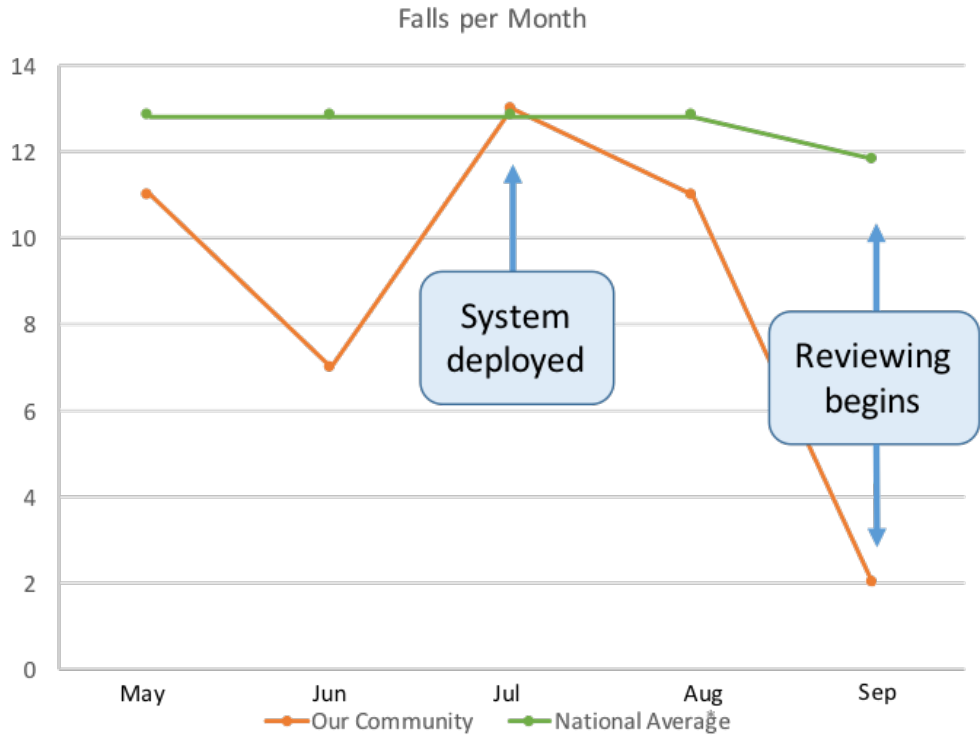


Figure 5.2: Fall rate. In the four months prior to video review, the fall rate at the community was 10.5 falls per month, 79% of the national average. In the final month, 2 falls occurred, 17% of the national average.

5.4 Results

During the three-month study period from July to September, 26 falls occurred. In the two months prior, 18 falls occurred. Overall, the rate was 10.5 falls per month before video review, and 2 falls occurred in the month following video review. The facility supported an average of 38 residents with a slight dip in the final month. For a facility with 38 residents, the expected fall rate is 12.7 falls per month based on the national average of 4 falls per month of individuals with dementia living in care facilities [17]. In the final month, the resident population declined slightly and thus, the national average is adjusted accordingly in Figure 2. The overall fall rate in this community was 79% of the national average for the 4 months prior to review and 17% of the national average in the month following fall review. Applying a one-tailed, two-sample t-test, the reduction in fall rate normalized by the number of residents cared for at the time is statistically significant with $p=0.030$.

5.5 Discussion

Although promising, these results are only preliminary. For example, in the final month 1 resident that fell four times in the previous month passed away. She fell zero times in the month before that. As can be seen in Figure 2, deviations in the fall risk naturally occur as some residents become greater risk before eventually passing away. If these 4 falls had been removed, there would only have been 7 falls in the second month, the same number of falls experienced in the month before the pilot study began. Controlling for this resident (i.e., removing her falls from all months), the result of the one-tailed two-sample t-test applied to the per resident fall rate drops to $p=0.058$, no longer significant at the often-used 0.05 threshold. Moreover, this t-test requires the assumption that the variance in the fall rate is the same before and after the introduction of video review. With only one sample data point, there is no evidence to support this assumption. Clearly, more data is needed validate that the fall rate in managed care facilities can be reduced through interactive video review of falls.

5.6 Conclusions

If verified, the impact of these results could be tremendous. Based on feedback from the families, the reduction from 10.5 falls per month to 2 falls in the final month is equivalent to approximately \$20k in savings in emergency room visits alone both for the families and for Medicare. More importantly, any one of these falls could have led to serious fracture, severe loss in mobility, significantly decreased quality of life, and significantly increased cost of care. Given that Alzheimers disease is the most expensive disease in the US, and fall accidents are the leading cause of hospitalization in Alzheimers care, this simple intervention may be the first steps toward a big impact. Based on this preliminary result, the operators of this care community have agreed to expand the study to 10-20 more facilities in California. In this next phase, we will deploy the same system, conduct video review after the first month, and observe if a statistically significant reduction in the fall rate occurs.

Chapter 6

Conclusions

6.1 Review of Project Report

In this project report we discuss 4 projects spanning a total of 2.5 years. We begin by discussing in Chapter 1 the relevant open problems in the dementia research community that could be supported by the development of new technologies. In the following 4 chapters we discuss several applications of machine learning and commercially available sensing to develop a new technologies for the Alzheimer's research community.

We first discuss the search for methods of curing, mitigating, or delaying the effects of Alzheimer's disease. Relevant research in this area includes monitoring the effects of diet, exercise, cognitive stimulation, and related factors on disease progression. To support this research area, in Chapter 2, we present the design and implementation of a wearable system called Max for collecting fine-grained measurements from environmental sensors and to perform analysis of this data for trends and anomalies. With this system, we hope to provide the Alzheimer's research community with the proper tools to perform functional monitoring of individual patients, to study how the effects of potential mitigating factors influence patients on an individual as opposed to a population level, and to perform new studies on how non-invasive home monitoring could be performed to recognize risks factors for conditions like urinary tract infections before they escalate into emergency room visits. The Max system is currently deployed with 18 individuals living in private homes and monitored by the the University of Nebraska Medical Center as part of the Dementia Care Ecosystem with data collection ongoing at the time this project report was submitted.

We next discuss the difficulty with accurately diagnosing Alzheimer's disease and related dementias. Since typical clinical diagnosis only provides accuracy near 75% for Alzheimer's disease, a focus here is on identifying biomarkers which are particularly indicative of a particular disease type. The ideal biomarker would be obtainable non-invasively and would identify both the disease and the stage of the disease. The most relevant biomarker for Alzheimer's disease involves measuring the concentration of molecules deriving from the characteristic amyloid- β in cerebral spinal fluid. In Chapter 3, we discuss another possible

biomarker which involves performing computational assessment of the speech patterns. We show that 92% accuracy can be achieved in matching an existing diagnosis for individuals with dementia based on one short conversation based on traditional machine learning approaches. Unfortunately, practical circumstances limit the scope of this study within the project report. Specifically, the much more interesting results require the collection of extensive longitudinal data which is not feasible within the scope of this project report – showing that the final clinical diagnosis performed from post-mortem histology can be determined with accuracies matching that of a human expert or that a screening tool for early diagnosis could be developed with high sensitivity based on a smartphone app alone.

We finally discuss how care could be improved for those currently living with Alzheimer’s disease and related dementias, paying particular attention to improving the quality of care and reducing the cost. This is a major research focus within the public health community since Alzheimer’s disease is currently the most expensive disease in the US and the number of affected individuals is continuing to rise at an ever-increasing rate [2]. In Chapters 4 and 5, we explore how existing methods in computer vision can be applied to fall detection and prevention. As discussed, since falls are the greatest cause of hospitalization in Alzheimer’s care and 3 in 4 fallers experience repeat falls [50], it seems that a technology capable of both detecting falls and showing caregivers how falls are occurring could provide significant benefits. In Chapter 4, we show that the technology is possible by applying existing domain adaptation techniques to a dataset of 200 falls acted out by healthy individuals. In this dataset, we show 92% precision and 86% recall in daylight recordings – comparable with results from wearable fall detection systems [6]. Given the deep-learning paradigm, the practical challenge in this case appears to arise from the shortage of data containing true falls collected with IR night-vision sensors. By applying the same techniques to a sufficiently large database of true falls, it appears technically feasible to perform fall detection with high accuracy from video. In Chapter 5, we study practical concerns with placing video cameras in memory care through one 3-month pilot study at a 40-resident memory care community in the San Francisco Bay Area. From active reviews of falls in the community, the management at this community reduced the fall rate by 80% during the study period. Moreover, possible concerns around the invasion of privacy for residents and staff appeared minimal given the possible safety benefits. Although preliminary, this result appears encouraging for reducing the fall rate in managed dementia care.

6.2 Final Conclusion: Hybrid Solutions are Required for Practical Challenges

From this work, the key conclusion is the need for hybrid solutions when solving practical problems given the current limitations of the machine learning methods discussed. Despite amazing successes in artificial intelligence, the best results continue to emerge from the application of supervised learning techniques to large datasets where clear loss functions can

be applied (e.g., image classification on ImageNet [16]). The need for large labeled training sets puts several limitations on the value which can be provided by fully automated systems. This is especially true for early stage companies that are continuing to learn about the most useful services they can provide and cannot afford the long design cycles and high resource costs required to develop large training sets.

The first type of hybrid solution is like that presented in Chapter 2 whereby a system like the Max smartwatch system must be designed before data collection can begin. In this case, we found that unsupervised anomaly detection methods provide the most value. Once an anomaly is detected, a human observer can step in to determine whether or not the anomaly actually signifies a noteworthy event. The difficulty in this case is designing a feature space whereby anomalous events are likely to contain information which is interesting to a human observer. For instance, in the functional monitoring case, we want to flag events where the affected individual is dramatically less active, but we do not want to flag every time she visits a new GPS location. Hence, anomaly detection algorithms are performed on the total Euclidean distance traveled in a given day, not on the raw GPS data itself.

The second type of hybrid solution is similar to the first but based on a semi-supervised instead of unsupervised approach. For instance, in Chapters 4 and 5, a system is discussed for detecting falls automatically, but evaluation of what caused the fall is left to a human observer. Thus, a more simple problem is posed which can fit within the current limitations of the computational methods, but reasoning about cause and effect is left to a human expert. This approach appears particularly fruitful in situations such as the fall detection problem where substantial benefit could be provided by a human expert reviewing all of the data, but without the use of computational detection, this review would not be practically feasible due to the volume of data collected. Another example of this is the screening tool for Alzheimer's disease based on speech discussed in Chapter 3.

The final type of hybrid solution is like that posed at the end of Chapter 3 whereby a highly accurate Alzheimer's detection system could be developed by implementing a system where human experts and trained computational models work side-by-side. In this type of system, the human is able to detect features which can be difficult for the computational model to discover such as difficulty responding to subtle social cues and the model is able to detect features which are difficult for the human to track such as changes in the frequency with which the affected individual uses the word 'you' over time. Although untested here, it would be very interesting to build a joint model to classify affected individuals based on the features and weights chosen both by a human expert and trained computational model.

Bibliography

- [1] National Institute on Aging: Alzheimer’s Disease Education and Referral Center. *Alzheimer’s Disease Fact Sheet*. Aug. 2016. URL: <https://www.nia.nih.gov/alzheimers/publication/alzheimers-disease-fact-sheet>.
- [2] Alzheimer’s-Association. “2016 Alzheimer’s Disease Facts and Figures”. In: *Alzheimer’s and Dementia* 11.3 (2016).
- [3] J. A. Anguera et al. “A Consensus on the Brain Training Industry from the Scientific Community”. In: *Stanford Center on Longevity* (Oct. 2014). URL: <http://longevity3.stanford.edu/blog/2014/10/15/the-consensus-on-the-brain-training-industry-from-the-scientific-community-2/>.
- [4] J. A. Anguera et al. “Video game training enhances cognitive control in older adults”. In: *Nature* 501 (Sept. 2013), pp. 97–101.
- [5] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. “Survey on speech emotion recognition : Features, classification schemes, and databases”. In: *Pattern Recognition* (Sept. 2010), pp. 572–587.
- [6] Fabio Bagala et al. “Evaluation of Accelerometer-Based Fall Detection Algorithms on Real-World Falls”. In: *PLoS ONE* 7.5 (2012).
- [7] David A. Balota et al. “The English Lexicon Project. 39”. In: *Behavior Research Methods*, 39 (2007), pp. 445–459.
- [8] L. J. Baraff et al. “Practice guideline for the ED management of falls in community-dwelling elderly persons”. In: *Annals of Emergency Medicine* 30 (1992), pp. 480–492.
- [9] Matthew Baumgart et al. “Summary of the evidence on modifiable risk factors for cognitive decline and dementia: A population-based perspective”. In: *Alzheimer’s and Dementia: The Journal of the Alzheimer’s Association* 11.6 (June 2015), pp. 718–726.
- [10] Pam Belluck. “Eli Lilly’s Experimental Alzheimer’s Drug Fails in Large Trial”. In: *The New York Times* (Nov. 2016). URL: <http://www.nytimes.com/2016/11/23/health/eli-lillys-experimental-alzheimers-drug-failed-in-large-trial.html>.
- [11] A. J. Blake, K. Morgan, and M. J. Bendall. “Falls by elderly persons at home: prevalence and associated factors”. In: *Age Ageing* 17 (1988), pp. 365–372.

- [12] Paul Boersma et al. “Praat, a system for doing phonetics by computer”. In: *Glott international* 5.9/10 (2002), pp. 341–345.
- [13] R.S. Bucks et al. “Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance”. In: *Aphasiology, vol. 14* (2000), pp. 2–35.
- [14] Maria Burke. “Why Alzheimer’s Drugs Keep Failing”. In: *Scientific American* (July 2014).
- [15] Julie P. W. Bynum et al. “The Relationship Between a Dementia Diagnosis, Chronic Illness, Medicare Expenditures, and Hospital Use”. In: *Journal of the American Geriatrics Society* 52.2 (2004), pp. 187–194.
- [16] Jia Deng et al. “ImageNet: A large-scale hierarchical image database”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, pp. 248–255.
- [17] Carol Van Doorn et al. “Dementia as a Risk Factor for Falls and Fall Injuries Among Nursing Home Residents”. In: *Journal of the American Geriatrics Society* 51.9 (2003), pp. 1213–1218.
- [18] Florian Eyben et al. “Recent developments in opensmile, the munich open-source multimedia feature extractor”. In: *Proceedings of the 21st ACM international conference on Multimedia*. ACM. 2013, pp. 835–838.
- [19] P. Feng et al. “Deep learning for posture analysis in fall detection”. In: *2014 19th International Conference on Digital Signal Processing*. Aug. 2014, pp. 12–17. DOI: 10.1109/ICDSP.2014.6900806.
- [20] Martin A. Fischler and Robert C. Bolles. “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography”. In: *Communications of the ACM* 24.6 (June 1981), pp. 381–395.
- [21] J. Fleming and C. Brayne. “Inability to Get up after Falling, Subsequent Time on Floor, and Summoning Help: Prospective Cohort Study in People over 90”. In: *BMJ* 337.1 (Nov. 2008).
- [22] Marshal F. Folstein, Susan E. Folstein, and Paul R. McHugh. “Mini-mental state: a practical method for grading the cognitive state of patients for the clinician”. In: *Journal of psychiatric research* 12.3 (1975), pp. 189–198.
- [23] Kathleen C. Fraser, Frank Rudzicz, and Elizabeth Rochon. “Using text and acoustic features to diagnose progressive aphasia and its subtypes”. In: *Proceedings of Interspeech* (2013).
- [24] Kathleen C. Fraser et al. “Automated classification of primary progressive aphasia subtypes from narrative speech transcripts”. In: *Cortex* (May 2014), pp. 43–60.
- [25] Kathleen Fraser et al. “Automatic speech recognition in the diagnosis of primary progressive aphasia”. In: *Fourth Workshop on Speech and Language Processing for Assistive Technologies* (2013), pp. 47–54.

- [26] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. “A Neural Algorithm of Artistic Style”. In: *arXiv:1508.06576 [cs, q-bio]* (Aug. 2015). arXiv: 1508.06576. URL: <http://arxiv.org/abs/1508.06576> (visited on 05/06/2016).
- [27] Dimitra G. Georganopoulou et al. “Nanoparticle-based detection in cerebral spinal fluid of a soluble pathogenic biomarker for Alzheimer’s disease”. In: *Proceedings of the National Academy of Sciences* 102.7 (2004), pp. 266–273.
- [28] Ross Girshick. “Fast R-CNN”. In: *arXiv:1504.08083 [cs]* (Apr. 2015). arXiv: 1504.08083. URL: <http://arxiv.org/abs/1504.08083> (visited on 05/09/2016).
- [29] Markus Goldstein and Seiichi Uchida. “A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data”. In: *PLOS ONE* (Apr. 2016).
- [30] Google. *Android Developer Guide*. 2016. URL: <https://source.android.com>.
- [31] ML Gorno-Tempini et al. “Classification of primary progressive aphasia and its variants”. In: *Neurology* 76.11 (2011), pp. 1006–1014.
- [32] Curry Guinn and Anthony Habash. “Language Analysis of Speakers with Dementia of the Alzheimer’s Type”. In: *Artificial Intelligence for Gerontechnology Technical Report* (2012), pp. 8–13.
- [33] Curry Guinn, Ben Singer, and Anthony Habash. “A Comparison of Syntax, Semantics, and Pragmatics in Spoken Language among Residents with Alzheimer’s Disease in Managed-Care Facilities”. In: *IEEE Symposium on Computational Intelligence in Healthcare and E-Health* (Dec. 2014).
- [34] P. J. Holliday et al. “Video recording of spontaneous falls of the elderly”. In: *American Society for Testing and Materials* (1990), pp. 7–16.
- [35] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. “Extreme learning machine: Theory and applications”. In: *Neurocomputing*, 70 (May 2006), pp. 489–501.
- [36] Stanislaw Jankowski et al. “Deep learning classifier for fall detection based on IR distance sensor data”. In: *IEEE*, Sept. 2015, pp. 723–727. ISBN: 978-1-4673-8359-2 978-1-4673-8361-5. DOI: 10.1109/IDAACS.2015.7341398. URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7341398> (visited on 05/09/2016).
- [37] Yangqing Jia et al. “Caffe: Convolutional Architecture for Fast Feature Embedding”. In: *arXiv preprint arXiv:1408.5093* (2014).
- [38] Keith A. Johnson et al. “Brain Imaging in Alzheimer Disease”. In: *Cold Spring Harb Perspect Med* 2.4 (Apr. 2012).
- [39] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python*. 2001–. URL: <http://www.scipy.org/>.

- [40] Yasuhiro Kakihara et al. “Acoustic Feature Selection Utilizing Multiple Kernel Learning for Classification of Children with Autism Spectrum and Typically Developing Children”. In: *Proceedings of the 2013 IEEE/SICE International Symposium on System Integration, Kobe International Conference Center, Kobe, Japan* (Dec. 2013), pp. 490–494.
- [41] Zaven S Khachaturian. “Revised criteria for diagnosis of Alzheimer’s disease: National Institute on Aging-Alzheimer’s Association diagnostic guidelines for Alzheimer’s disease”. In: *Alzheimer’s & dementia: the journal of the Alzheimer’s Association* 7.3 (2011), pp. 253–256.
- [42] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 1097–1105. URL: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>.
- [43] Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. “Age of acquisition ratings for 30,000 English words”. In: *Behavioral Research* (May 2012), pp. 978–990.
- [44] Robert W Levenson and John M Gottman. “Marital interaction: physiological linkage and affective exchange.” In: *Journal of personality and social psychology* 45.3 (1983), p. 587.
- [45] Weifeng Liu, Puskal P. Pokharel, and Jose C. Principe. “The Kernel Least-Mean-Square Algorithm”. In: *IEEE Transactions on Signal Processing IEEE Trans. Signal Process.* 56.2 (2008), pp. 543–554.
- [46] Hong Lu et al. “Intelligent Human Fall Detection for Home Surveillance”. In: IEEE, Dec. 2014, pp. 672–676. ISBN: 978-1-4799-7646-1. DOI: 10.1109/UIC-ATC-ScalCom.2014.56. URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7307023> (visited on 05/09/2016).
- [47] Bruce L. Miller and Bradley F. Boeve. *The Behavioral Neurology of Dementia*. Cambridge University Press, 2009.
- [48] David Neary et al. “Frontotemporal lobar degeneration A consensus on clinical diagnostic criteria”. In: *Neurology* 51.6 (1998), pp. 1546–1554.
- [49] *Neural Artistic Style in Python*. {https://github.com/andersbll/neural_artistic_style}. URL: https://github.com/andersbll/neural_artistic_style (visited on 05/10/2016).
- [50] M. C. Nevitt et al. “Risk factors for recurrent nonsyncopal falls: a prospective study”. In: *JAMA* 261 (1989), pp. 1663–2668.
- [51] Serguei V. S. Pakhomov et al. “Computerized Analysis of Speech and Language to Identify Psycholinguistic Correlates of Frontotemporal Lobar Degeneration”. In: *Cognitive and Behavioral Neurology* (Sept. 2010), pp. 165–177.

- [52] Anindya S. Paul and Eric A. Wan. “RSSI-Based Indoor Localization and Tracking Using Sigma-Point Kalman Smoothers”. In: *IEEE Journal of Selected Topics in Signal Processing* 3.5 (Oct. 2009), pp. 543–554.
- [53] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [54] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [55] Andrew Pollack. “New Data on 2 Alzheimer’s Drugs Alters Hope and Expectation”. In: *The New York Times* (July 2015). URL: <http://www.nytimes.com/2015/07/23/business/new-data-on-2-alzheimers-drugs-alters-hope-and-expectation.html>.
- [56] Katherine L. Possin and Bruce L. Miller. *Care Ecosystem: Navigating Patients and Families Through Stages of Care*. 2016. URL: <https://www.nia.nih.gov/alzheimers/clinical-trials/care-ecosystem-navigating-patients-and-families-through-stages-care>.
- [57] Peter S. Pressman and Gorno-Tempini. “Introduction and History of Primary Progressive Aphasia”. In: *Neurobiology of Language* (2015). Ed. by Greg Hickock and Steve Small.
- [58] Peter S Pressman and Bruce L Miller. “Diagnosis and management of behavioral variant frontotemporal dementia”. In: *Biological psychiatry* 75.7 (2014), pp. 574–581.
- [59] D. Prudham and J. G. Evans. “Factors associated with falls in the elderly: a community study”. In: *Age Ageing* 10 (1981), pp. 141–146.
- [60] George W. Rebok et al. “Ten-Year Effects of the Advanced Cognitive Training for Independent and Vital Elderly Cognitive Training Trial on Cognition and Everyday Functioning in Older Adults”. In: *Journal of the American Geriatrics Society* 62.1 (Jan. 2014), pp. 16–24.
- [61] Benjamin Recht. *UC Berkeley Course Notes for EE227C: Convex Algorithms*. 2015.
- [62] Steven N. Robinovitch et al. “Video Capture of the Circumstances of Falls in Elderly People Residing in Long-term Care: An Observational Study”. In: *The Lancet* 381.9860 (2013), pp. 47–53.
- [63] Laurence Z. Rubenstein et al. “The Value of Assessing Falls in an Elderly Population”. In: *Annals of Internal Medicine* 113.4 (1990), p. 308.
- [64] Frank Rudzicz et al. “Automatically Identifying Trouble-indicating Speech Behaviors in Alzheimer’s Disease”. In: *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*. ASSETS ’14. Rochester, New York, USA: ACM, 2014, pp. 241–242. ISBN: 978-1-4503-2720-6. DOI: 10.1145/2661334.2661382. URL: <http://doi.acm.org/10.1145/2661334.2661382>.

- [65] Gerard Salton and Christopher Buckley. “Term-weighting approaches in automatic text retrieval”. In: *Information processing & management* 24.5 (1988), pp. 513–523.
- [66] Bjorn Schuller and Anton Batliner. *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013. Chap. 8.
- [67] Leslie M. Shaw et al. “Cerebrospinal fluid biomarker signature in Alzheimer’s disease neuroimaging initiative subjects”. In: *Annals of Neurology* 65.4 (Apr. 2009), pp. 403–413.
- [68] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *arXiv:1409.1556 [cs]* (Sept. 2014). arXiv: 1409.1556. URL: <http://arxiv.org/abs/1409.1556> (visited on 05/06/2016).
- [69] Audacity Team. *Audacity (Version 2.0.5)*. 2014. URL: <http://audacity.sourceforge.net/>.
- [70] Calvin Thomas et al. “Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech”. In: *IEEE International Conference Mechatronics and Automation, 2005*. Vol. 3. IEEE. 2005, pp. 1569–1574.
- [71] Mary E. Tinetti, Wen-Liang Liu, and Elizabeth B. Claus. “Predictors and Prognosis of Inability to Get Up After Falls Among Elderly Persons”. In: *JAMA* 261.1 (1993), p. 65.
- [72] Mary E. Tinetti, M. Speechley, and S. F. Ginter. “Risk factors for falls among elderly persons living in the community”. In: *N Engl J Med* 319 (1988), pp. 1701–1707.
- [73] Mary E. Tinetti, T. Franklin Williams, and Raymond Mayewski. “Fall Risk Index for Elderly Patients Based on Number of Chronic Disabilities”. In: *The American Journal of Medicine* 80.3 (1986), pp. 429–434.
- [74] Eric Tzeng et al. “Simultaneous Deep Transfer Across Domains and Tasks”. In: IEEE, Dec. 2015, pp. 4068–4076. ISBN: 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.463. URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7410820> (visited on 05/06/2016).
- [75] Carl Vondrick, Donald Patterson, and Deva Ramanan. “Efficiently Scaling up Crowdsourced Video Annotation”. In: *International Journal of Computer Vision* (). 10.1007/s11263-012-0564-1, pp. 1–21. ISSN: 0920-5691. URL: <http://dx.doi.org/10.1007/s11263-012-0564-1>.
- [76] Kevin Weekly et al. “Indoor Occupant Positioning System Using Active RFID Deployment and Particle Filters”. In: *DCOSS ’14 Proceedings of the 2014 IEEE International Conference on Distributed Computing in Sensor Systems*. May 2014, pp. 35–42.
- [77] D. Wild, U. S. Nayak, and B. Isaacs. “How Dangerous Are Falls in Old People at Home?” In: *BMJ* 282.6260 (1981), pp. 266–268.

- [78] Josh D Woolley et al. “The diagnostic challenge of psychiatric symptoms in neurodegenerative disease: rates of and risk factors for prior psychiatric diagnosis in patients with early neurodegenerative disease”. In: *The Journal of clinical psychiatry* 72.2 (2011), pp. 126–133.
- [79] R. Woolrych et al. “Exploring the potential of using real life video capture to investigate the circumstances of falls among older adults in long-term care”. In: *Gerontechnology* 13.2 (2014), pp. 132–133.
- [80] M. Yu et al. “A Posture Recognition-Based Fall Detection System for Monitoring an Elderly Person in a Smart Home Environment”. In: *IEEE Transactions on Information Technology in Biomedicine* 16.6 (Nov. 2012), pp. 1274–1286. ISSN: 1089-7771. DOI: 10.1109/TITB.2012.2214786.
- [81] Anthony Zhang. *Speech Recognition (Version 2.0) [Software]*. 2015.
- [82] Han Zou et al. “An RFID indoor positioning system by using weighted path loss and extreme learning machine”. In: *Cyber-Physical Systems, Networks, and Applications (CPSNA), 2013 IEEE 1st International Conference on*. Aug. 2013.