ADVANCING MOBILE SENSING IN DYNAMIC ENVIRONMENTS

Weichen Wang
Dartmouth College, weichen.wang.gr@dartmouth.edu

Follow this and additional works at: https://digitalcommons.dartmouth.edu/dissertations

Part of the Computer Sciences Commons

Recommended Citation
Wang, Weichen, "ADVANCING MOBILE SENSING IN DYNAMIC ENVIRONMENTS" (2024). Dartmouth College Ph.D Dissertations. 254.
https://digitalcommons.dartmouth.edu/dissertations/254

This Thesis (Ph.D.) is brought to you for free and open access by the Theses and Dissertations at Dartmouth Digital Commons. It has been accepted for inclusion in Dartmouth College Ph.D Dissertations by an authorized administrator of Dartmouth Digital Commons. For more information, please contact dartmouthdigitalcommons@groups.dartmouth.edu.
ADVANCING MOBILE SENSING IN DYNAMIC ENVIRONMENTS

A Thesis
Submitted to the Faculty
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy

in
Computer Science

by Weichen Wang

Guarini School of Graduate and Advanced Studies
Dartmouth College
Hanover, New Hampshire

April 2024

Examinining Committee:

Andrew T. Campbell, Ph.D., Chair

Yaoqing Yang, Ph.D.

Nicholas C. Jacobson, Ph.D.

Anind Dey, Ph.D.

F. Jon Kull, Ph.D.
Dean of the Guarini School of Graduate and Advanced Studies
Abstract

This thesis presents a comprehensive exploration of enhancing mobile sensing capabilities to address various aspects of human behavior, mental health, personality, social functioning and beyond. We redesign the StudentLife app to improve its sensing efficiency and dependability, enabling support for multi-year-long studies. By adopting new app design, this study addresses the technical challenges of continuous sensing and enhances system robustness. The work is organized into several key studies that collectively aim to expand the scope of mobile sensing in diverse and complex environments.

The first study broadens the scope of mobile sensing to assess personality traits, exploring the potential of within-person variability in behavior to predict personality traits. By analyzing data from 646 college students, this study demonstrates significant correlations between sensed behaviors and self-reported personality traits, offering a novel approach for passive personality assessment.

The second study utilizes mobile sensing data to provide insights into the social functioning of individuals with mental health disorders, specifically schizophrenia. This study identifies behavioral patterns correlated with various aspects of social functioning, highlighting the potential for mobile sensing to inform new assessment and intervention strategies.
The third study investigates the integration of voice diaries to enhance the prediction of auditory verbal hallucination (AVH) severity. This approach leverages deep learning models to analyze speech and mobility data, showcasing the feasibility of using mobile sensing for in-the-wild psychiatric symptom assessment.

The fourth study predicts the mental well-being of college students with a special emphasis on first-generation students, using longitudinal mobile sensing data to identify risk factors and behavioral patterns associated with mental health.

Finally, the thesis investigates the challenge of domain drift and model degradation over time, exploring adaptation technologies to maintain the effectiveness of mobile sensing frameworks for depression detection. By analyzing passive sensing data and self-reported surveys from undergraduate students over several years, this work demonstrates the efficacy of domain adaptation strategies in ensuring robust depression detection.

Together, these studies contribute to the development of power-efficient, scalable, and adaptable mobile sensing systems, pushing the boundaries of mobile sensing technologies, and offering new perspectives on assessing mental health and beyond.
Acknowledgments

Firstly, I would like to express my gratitude to my advisor, Prof. Andrew T. Campbell, for his unwavering support throughout my Ph.D. studies and research. His influence has significantly shaped my approach to research, emphasizing the creation of impactful work. He fostered an environment that encouraged collaboration with other scholars, for which I am deeply appreciative. His honest feedback, patience, and understanding have been invaluable, and without his guidance and support, completing this thesis would not have been possible.

Special thanks to Prof. Gabriella Harari, Prof. Xia Zhou, Prof. Dror Ben-Zeev, Dr. Rui Wang and Dr. Jeremy F Huckins for their guidance and support at various phases of my Ph.D. journey.

I am also grateful for the opportunity to work closely with Prof. Benjamin Buck, Prof. Trevor Cohen, Prof. Meghan Meyer, Prof. Tanzeem Choudhury, Prof. Akane Sano, Dr. Sandrine R. Müller, Dr. Xuhai Xu, Dr. Alex W DaSilva, Weizhe Xu, Rachel Brian, Ayesha Chander, and Courtney Rogers.

I am also thankful to all the current and former members of the Dartnets Lab: Rui Wang, Tianxing Li, Zhao Tian, Shayan Mirjafari, Kizito Masaba, Yichen Li, Mikio Obuchi, Subigya Nepal, Ruibo Liu, Qijia Shao, and Arvind Pillai for their camaraderie and the time spent together.
Lastly, I am deeply thankful for my family’s support. A special thank you to my wife, Jialing Wu, for her love, understanding, and encouragement. And to my parents, whose endless support and provision of a solid educational foundation have been cornerstones of my journey.
Contents

Abstract ................................................................. ii
Acknowledgments .................................................... iv

1 Introduction  .......................................................... 1
  1.1 Overview ......................................................... 1
  1.2 Methodology: Smartphone Sensing Systems and More ........... 4
    1.2.1 Overview of the StudentLife Sensing Systems ............... 4
    1.2.2 Key Updates on the StudentLife Sensing Systems ....... 6
    1.2.3 Feature Extraction ....................................... 8
    1.2.4 Modeling Methods ....................................... 11
    1.2.5 Ground Truth ........................................... 12
  1.3 Problem Statement ............................................. 13
    1.3.1 Broadening the Scope of Mobile Sensing for Personality Assessment ......................................................... 13
    1.3.2 Utilizing Mobile Sensing for Social Functioning Analysis .... 14
    1.3.3 Enhancing Auditory Verbal Hallucination (AVH) Severity Prediction through Extended Sensing Signals ............ 15
1.3.4 Predicting Mental Well-being of College Students with a Focus on First-Generation Students ........................................ 16
1.3.5 Addressing Domain Drift and Model Degradation through Adaptation Technologies ..................................................... 17
1.4 Protection of Human Subjects ......................................................... 18

2 Sensing Behavioral Change over Time: Using Within-Person Variability Features from Mobile Sensing to Predict Personality Traits 20
2.1 Introduction ................................................................................. 20
2.2 Related Work .............................................................................. 24
2.3 Within-person Variability Measurements ........................................ 26
2.3.1 Behaviors Inferred from Passive Smartphone Sensing .............. 27
2.3.2 Within-Person Variability Metrics .............................................. 29
2.4 Data Collection and Processing .................................................... 36
2.4.1 Data Inclusion Criteria .............................................................. 37
2.4.2 Big Five Personality Ground truth ............................................ 38
2.5 Assessing Personality using Within-person Variability Measures .......... 40
2.5.1 Association Analysis ................................................................. 40
2.5.2 Prediction Analysis ................................................................. 46
2.6 Limitations ................................................................................. 51
2.7 Conclusion and Future Work ....................................................... 55

3 Social Sensing: Assessing Social Functioning of Patients Living with Schizophrenia using Mobile Phone Sensing 58
3.1 Introduction ................................................................................. 58
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.1</td>
<td>Procedures and Participants</td>
<td>100</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Ground Truth</td>
<td>101</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Dataset</td>
<td>103</td>
</tr>
<tr>
<td>4.4</td>
<td>Data Preprocessing</td>
<td>106</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Interpretable Feature Engineering for Traditional Machine Learning</td>
<td>107</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Data Preprocessing for Deep Learning</td>
<td>112</td>
</tr>
<tr>
<td>4.5</td>
<td>Predicting Models and Results</td>
<td>115</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Baseline: Using Traditional Machine Learning Approaches</td>
<td>116</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Uni-modal Data</td>
<td>117</td>
</tr>
<tr>
<td>4.5.3</td>
<td>Multi-modal Data</td>
<td>120</td>
</tr>
<tr>
<td>4.6</td>
<td>Interpretation</td>
<td>121</td>
</tr>
<tr>
<td>4.6.1</td>
<td>Important Interpretable Features</td>
<td>122</td>
</tr>
<tr>
<td>4.6.2</td>
<td>Interpretation of Auto-Learned Features from Deep Learning</td>
<td>126</td>
</tr>
<tr>
<td>4.6.3</td>
<td>Learned Attention Vector in Sentence Embedding</td>
<td>134</td>
</tr>
<tr>
<td>4.7</td>
<td>Discussion</td>
<td>134</td>
</tr>
<tr>
<td>4.7.1</td>
<td>Using Multimodal Streams and Deep Learning to Predict AVH</td>
<td>134</td>
</tr>
<tr>
<td>4.7.2</td>
<td>Fully Automatic Pipeline (Using Automated Speech Recognition)</td>
<td>136</td>
</tr>
<tr>
<td>4.7.3</td>
<td>Implications</td>
<td>137</td>
</tr>
<tr>
<td>4.8</td>
<td>Conclusion</td>
<td>138</td>
</tr>
</tbody>
</table>

5 First-Gen Lens: Assessing Mental Health of First-Generation Students across Their First Year at College Using Mobile Sensing

5.1 Introduction                                                                 | 139  |
5.2 Related Work                                                              | 146  |
5.3 First Year in a Nutshell .................................................. 147
  5.3.1 College Life ......................................................... 148
  5.3.2 First-Year Data ..................................................... 149
5.4 First-Gen Methodology .................................................. 152
  5.4.1 Study Design ....................................................... 152
  5.4.2 High School Survey ............................................... 153
  5.4.3 StudentLife: Behavioral Sensing App ......................... 153
  5.4.4 Ecological Momentary Assessment (EMA): Self-reported Men-
       tal Health ............................................................ 158
5.5 Analysis ........................................................................ 158
  5.5.1 Q1: How Do the Risk Factors Associated with First-generation
       Students Relate to Their Mental Health during Their First Year
       at College? ............................................................... 159
  5.5.2 Q2: What are the Key Behavioral Differences between First-
       generation and Non-first-generation Students across Each Term
       as They Progress through Their First Year? ...................... 164
  5.5.3 Q3: What Behaviors Do First-generation Students Exhibit That
       Are Associated with Better Mental Health? And Are These Be-
       haviors Similar to Those Exhibited by the Non-first-generation
       Students? ............................................................... 168
  5.5.4 Q4: How Can We Predict the Mental Well-being of At-risk
       Students More Accurately, with a Particular Focus on the First-
       generation Students? ............................................... 172
5.6 Discussion ................................................................. 182
5.6.1 Contribution of This Work ........................................ 183
5.6.2 The Benefit of Our Deep Learning Approach .................. 185
5.6.3 College Support for First-generation Students ................. 186
5.6.4 Limitations .......................................................... 189
5.7 Conclusion ............................................................... 191

6 The Need to Adapt: Mitigating Model Degradation in Longitudinal
Human Behavioral Sensing Studies 192
6.1 Introduction .............................................................. 192
6.2 Related Work .......................................................... 194
   6.2.1 Mobile Sensing and Mental Health .............................. 194
   6.2.2 Domain Adaptation .............................................. 195
6.3 Dataset ................................................................. 196
   6.3.1 Data Collection .................................................. 197
   6.3.2 Passive Sensing Data ............................................. 197
   6.3.3 Ground Truth .................................................... 198
6.4 Analysing the Domain Drift ........................................... 199
   6.4.1 Data Drift ....................................................... 200
   6.4.2 Concept Drift ................................................... 206
6.5 Detecting Model Degradation ......................................... 210
   6.5.1 Training Details ................................................. 212
   6.5.2 Result Analysis .................................................. 213
6.6 Mitigating Model Degradation using Domain Adaptation Algorithms 215
   6.6.1 Unsupervised Domain Adaptation .............................. 215
   6.6.2 Result Analysis .................................................. 217
List of Tables

2.1 Description of the features computed .................................. 34
2.2 Cost definition in FRI calculation ........................................ 36
2.3 Distributions of the personality ground truth .......................... 39
2.4 Big Five traits in relation to regularity using mixed-effect model. . 45
2.5 Prediction performance ..................................................... 50

3.1 Demographics of participants ............................................. 67
3.2 Statistics of Social Functioning Scale .................................. 68
3.3 Features computed from mobile sensing data. We have 115 behavioral features in total. The length of each feature is 160, corresponding to the number of SFS responses adopted in our dataset. The unit could be in seconds (e.g., duration of various activities, conversations, etc.), an integer of counts (number of places visited, number of lock/unlocks, etc.), or ambient light level/ambient sound amplitude directly from sensors. ................................................................. 70
3.4 Patterns associated with 7 sub-areas of social functioning .......... 76
3.5 Prediction Summary using leave-one-subject-out and customized leave-one-subject-out ................................................. 81
4.1 Demographics of the participants. ................................................. 102
4.2 Linguistic features ................................................................. 111
4.3 Predictive performance on AVH using uni-model data. ............... 120
4.4 Predictive performance on AVH using multi-model data. .......... 121
4.5 Confusion matrix: using architecture (Fig. 4.7b) and the "audio +
text" model (weighted f-1 score = 0.78) ................................. 123
4.6 Confusion matrix: using architecture (Fig. 4.7b) and the "combined-
all" model (weighted f-1 score = 0.84) ................................. 123
4.7 Top-3 representations automatically learned from audio data of voice
diaries that identify patients with severe AVH, as well as the top-
3 handcrafted acoustic features with the strongest correlation to the
representations. ................................................................. 129
4.8 Top-3 representations automatically learned from text data from voice
diaries that differentiate patients with severe AVH the most effectively,
as well as the top-3 created linguistic features with the strongest cor-
relation to the representations. ............................................. 130
4.9 Top-3 hand-crafted features with the strongest correlation to the auto-
learned representations from mobile sensing data. .................... 133
4.10 Predictive performance on AVH using multi-model data (with text data
from ASR). See Fig. 4.7 for the definition of structure (a) and (b). . . 137

5.1 High School Survey based on the CIRP Freshmen Survey (TFS) [283]:
Questions associated with risk factors. ................................... 154
5.2 Compliance in existing literature. ............................................ 156
5.3 Features computed from StudentLife mobile sensing data. Daily and epoch-based sensing features are computed across meaningful daily epochs: night/morning (12am-9am), day (9am-6pm), and evening (6pm-12am), allowing analysis of behavioral features and trends across different periods of the 24-hour clock).

5.4 Comparison of risk factors between first-generation and non-first-generation students using T-test. The scores of risk factors are normally distributed (p > 0.05 using KS-test). Full statistical results on subset A and subset B can be found in the supplementary document.

5.5 Comparison of PHQ-4 scores between first-generation and non-first-generation students using mixed effects with repeated PHQ-4 scores nested within the individual, and the group contrast (1: 1st-gens & 0: non-first-gens) treated as a fixed effect. First-generation students had significantly higher depression scores during the fall term only. The coefficients $\beta$ decrease as the year progresses, possibly indicating that first-generation students adapt to overcome initial risk factors. Full statistical results on subset A and subset B can be found in the supplementary document.

5.6 Associations between risk factors and mental health using mixed effects models with the repeated PHQ-4 nested within the individual, and each risk factor (Likert scale from 1 to 5) treated as a fixed effect. Better lifestyle and sociability are associated with better mental health during the first year at college (i.e., lower PHQ-4 score). Full statistical results on subset A and subset B are available in the supplementary document.
5.7 Comparison of behavioral features between first-generation and non-first-generation students. Behavioral sensing data and inferences are computed daily. They are analyzed using mixed effects models nested within individual and the first-gens/non-first-gens group contrast (1: 1st-gens & -1: non-first-gens) treated as the fixed effect. Associations between behavioral features and risk factors are examined using mixed effects models with the daily behavioral features nested within the individual. Each risk factor (Likert scale from 1 to 5) is treated as a fixed effect. The statistical analysis for the fall term is based on data collected after midterm, when participants were completely enrolled. Full statistical results on two subsets are available in the supplementary document.  

5.8 The coefficients of LASSO (Least Absolute Shrinkage and Selection Operator) regression for estimating PHQ-4.  

5.9 Predictive performance on PHQ-4 using traditional machine learning approaches. Traditional machine learning approaches fail to make predictions among first generations. Metrics are weighted to account for label imbalance.  

5.10 The details of DNN-1st-gen, DNN-term and DNN-PHQ in the architecture.  

5.11 Predictive performance of depression (PHQ-4) when different weights are assigned to the main loss and the auxiliary losses. Metrics are weighted to account for label imbalance.
6.1 Features: The table below lists the passive sensing data we collect from the participants. Daily and epoch-based sensing features are computed across meaningful daily epochs: night/morning (12am-9am), day (9am-6pm), and evening (6pm-12am), allowing analysis of behavioral features and trends across different periods of the 24-hour clock).

6.2 Evaluation of Data Drift During Baseline, Covid, and New-Norm Terms. Mixed Effects Models, nested within individuals, assess the variation, with contrasts between test groups (Covid terms / new-norm terms) and the control group (Baseline terms) designated as the fixed effect (1 for test groups, 0 for the control group). Significance levels are adjusted for multiple comparisons using the Benjamini/Hochberg False Discovery Rate (FDR) correction, with the following notations: (***, \( p < .01 \), **, \( .01 \leq p < .05 \), *, \( .05 \leq p \leq .10 \)). Values shaded in gray within the table indicate non-significant results, denoting no apparent change; they are included for informational purposes only.
6.3 Evaluation of Concept Drift (the association between behavior and depression) during Baseline, Covid, and New-Norm Terms. The student population was categorized into two groups: the non-depressed as the control group and the depressed as the testing group. Mixed Effects Models, nested within individuals, assess the variation, with contrasts between the test group (students with depression) and the control group (students without depression) designated as the fixed effect (1 for test groups, 0 for the control group). Significance levels are adjusted for multiple comparisons using the Benjamini/Hochberg False Discovery Rate (FDR) correction, with the following notations: (**p < .01, * .01 ≤ p < .05, * .05 ≤ p ≤ .10). Values that are not shaded gray and lack notations had p < .05 before FDR but do not maintain significance after the FDR correction was applied. Values shaded in gray within the table indicate non-significant results, denoting no apparent change; they are included for informational purposes only.

6.4 Source and target dataset

6.5 Data-point weight adjustment based on PHQ-2 scores and symptom lasting length when aggregating weekly self reports to a term
List of Figures

1.1 The StudentLife dashboard used to “datasit” the study. ............... 8

2.1 Circadian Rhythm analysis using audio amplitude sensing data from two participants. ......................... 32

2.2 System and study design. ........................................ 38

2.3 Histograms of the Big Five scores. The X axis displays the value of the score, which ranges from 1 (lowest) to 5 (highest). The Y axis shows the number of individuals that fall into the specific score bins. The three rows show the distribution of the scores for all participants (first row), the Android (second row) and the iOS (third row) users. ......... 39

2.4 Predicted values of personality traits and ground truth. .............. 52

3.1 The cumulative variability in the data explained by the top-n principal components. We include the top 23 principal components that explain 90% of the variance in our analysis. ................................. 72
3.2 Case study: the dynamics of a participant’s assessed and predicted social functioning using CLOSO. (a) the model identifies the strengths and weaknesses between the different components in social functioning. (b) the model tracks the dynamics of social functioning.

4.1 User interface of the daily voice diary in study app. (A) Welcoming page (B) Recording page.

4.2 We received 4809 audio diaries in all. (a) The majority of the voice diaries were recorded between the hours of 9 a.m. and 9 p.m., when the voice EMA randomly prompts. Others were recorded before or after the typical EMA time. (b) Many voice diaries are less than 30 seconds and contain simply random noise or a few words that are insufficient for analysis. There are also a significant number of audio diaries ranging in length from 2.5 minutes to 3 minutes (the maximum length allowed by the application).

4.3 The word cloud of audio diaries among people with non-severe and severe AVH.

4.4 Mel Spectrogram computed from one of our voice diaries.

4.5 Illustration of Bidirectional Gated Recurrent Unit network.

4.6 Illustration of self-attentive BiGRU network.

4.7 Illustration of deep neural network using multi-modal data (audio, text and sensor data).

4.8 SHAP value summary plot of the XGBoost model trained on hand-crafted features.
4.9 The distribution of the most discriminative representations from audio
data among the non-severe (label 0) and severe (label 1) AVH group.
The effect size is indicated by partial eta squared. . . . . . . . . . . . 128

4.10 The distribution of the most discriminative transcript representations
between the non-severe (label 0) and severe (label 1) AVH groups,
effect size indicated by partial eta squared. . . . . . . . . . . . . . . . 131

4.11 The distribution of the most discriminative sensing data representa-
tions between the non-severe (label 0) and severe (label 1) AVH groups.132

4.12 Attention vector of a segment of an audio diary. . . . . . . . . . . . . 133

5.1 Mental health dynamics, sleep and physical activity over the first year.
Plots are based on the mean value and 25/75 percentiles of each mea-
surement among students, grouped by first and non-first generations.
Participants are completely enrolled by mid-October. The plot in-
dicates that the first half of the fall term (Sep-Oct) has insufficient
participation. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 150

5.2 Flow of analysis of research questions addressed by the First-Gen study 159

5.3 Adding two auxiliary output in a neural network architecture where
the losses are provided by the labels of whether the student is a first-
generation student and the academic/college life progress. . . . . . . . 176

5.4 Top 10 most important features from deep learning model discovered
using permutation feature importance algorithm. The x-axis displays
the decrease in the F1-score when a feature is permuted. The y-axis
displays the name and index of each of the 166 features. . . . . . . . 181
5.5 The attention vector indicates the weights of different days during a term in predicting depression (PHQ-4) from sensing data from a person.

6.1 Model performance degradation of predicting depression in term when applying a model trained on source data (orange) on various target data (blue).

6.2 Model performance of using DANN (blue) and MCD (green) domain adaptation algorithms on target dataset. The gray bar is the performance without using any adaptation algorithms and is included for comparison purpose.

6.3 Model performance of using semi-supervised DANN (blue) and MCD (green) domain adaptation algorithms with anchor labels on target dataset.
Chapter 1

Introduction

1.1 Overview

Mobile sensing leverages the sensors embedded in mobile devices, such as smartphones, to unobtrusively collect data related to user behavior and environmental factors over time. This approach enables the capture of a wide array of data types including physical activity, social interactions, location, phone usage patterns, and ambient environmental factors like noise levels. Mobile sensing is especially useful in behavioral and mental health research for its ability to provide high-resolution, continuous streams of data without requiring active user input, thus minimizing burden and potential biases associated with self-report methods.

Back in 2013, researchers at Dartmouth College built the StudentLife [342] smartphone app and sensing system developed to automatically infer human behavior, aimed specifically at the college student population. The StudentLife app integrates Mobile Ecological Momentary Assessment (EMA) components to probe students’ states, such as stress and mood, across the term. The study conducted with Stu-
1.1 Overview

StudentLife was among the first to use automatic and continuous smartphone sensing to assess mental health, academic performance, and behavioral trends of a student body. It demonstrated strong correlations between automatically sensed data (such as conversation activity, mobility, and sleep) and a broad set of well-known mental well-being measures, including PHQ-9 depression, perceived stress (PSS), flourishing, and loneliness scales. Furthermore, it proposed a model to predict students’ cumulative GPA using automatic behavioral sensing data from smartphones, marking a significant advancement in the use of technology for monitoring and understanding student life and well-being. The results of the StudentLife study provide insights into the potential of mobile technology to support mental health and academic performance assessment and intervention in the student population.

Since then, mental health sensing is increasingly using mobile devices and wearables for passive, real-time monitoring of individuals in various settings [240]. Presently, advanced sensors in mobile phones and wearable devices facilitate the unobtrusive collection of a wide array of behavioral data without requiring active participation from users. Several studies have established connections between data collected through mobile sensing and various aspects of mental health, including depression [342, 347, 53, 287], anxiety [162, 286, 38], and mood fluctuations [203, 374]. Previous research predominantly had been focusing on mental health over short periods and often with a limited number of participants. In contrast, the thesis pushes the boundaries of this field in several significant ways.

First, we embark on a novel exploration into the realm of personality assessment within a sizable cohort of 646 students. This pioneering study not only expands the population size but also delves into a relatively uncharted area of research by assessing
1.1 Overview

personality traits through mobile sensing data.

Second, we extend our research scope to include social functioning among individuals with mental health disorders. By harnessing mobile sensing technology, we provide a more detailed and objective analysis of how effectively such technology can measure and predict social interactions and behaviors.

Third, we study Auditory Verbal Hallucinations (AVHs), a complex set of experiences that occur in a variety of forms across a wide range of people, including people with a variety of mental health diagnoses as well as by people who have no identifiable psychiatric or neurological diseases. We enrich mobile sensing with in-the-wild voice diaries to understanding auditory verbal hallucinations (AVH). We collect 4809 audio diaries from N=384 subjects over a one-month-long study period. Moreover, our ambitious initiative enrolled 400 participants across the United States, marking a substantial increase in scale and diversity compared to previous studies.

Fourth, we collected passive sensing data and self-reported surveys from 215 undergraduate student at Dartmouth College, spanning from September 2017 to June 2022. closely observe a group of students throughout their entire first year of college life. We focus on more complex and nuanced aspects of the study such as attention to minority groups like first-generation students. This longitudinal approach allows us to capture the evolution of behaviors and stressors as students transition into university life, a critical phase that has been seldom scrutinized in such depth.

Lastly, building on the fourth study, we continue to track these students over the entirety of their four-year college journey. This long-term analysis is crucial for understanding the domain drifts—how individuals’ behaviors and mental health indicators change over time—and for deploying adaptation technologies to counteract
1.2 Methodology: Smartphone Sensing Systems and More

the degradation of model accuracy due to these changes.

By integrating these five areas of research, we aim to create a comprehensive sensing framework that not only captures a wide array of behavioral data but also addresses the challenges of longitudinal studies. Our work promises to enhance the robustness of predictive models, adapt to the shifts in the data domain over extended periods, and refine our understanding of the complexities of human behavior.

1.2 Methodology: Smartphone Sensing Systems and More

In this subsection, we review the methodology underpinning our research, including our smartphone sensing systems, our approaches to data processing, the modeling methods employed, and the outcomes we aim to achieve. This overview provides a foundational understanding of how we collect, process, and analyze mobile sensing data to investigate behavioral patterns, and predict outcomes.

1.2.1 Overview of the StudentLife Sensing Systems

A smartphone sensing system typically consists of two main components: a sensing application on the user’s device and a backend server hosted in the cloud. The sensing app on the smartphone functions by collecting data from an array of phone sensors. This data is then packaged and uploaded to the cloud-based backend server. On the server side, there are several responsibilities: presenting web pages for study enrollment, managing data persistence, extracting behavioral features from the collected data, and providing visualization tools for administrators to track and monitor study
1.2 Methodology: Smartphone Sensing Systems and More

compliance.

The StudentLife sensing system[342], initially developed in 2013, has undergone continual updates to accommodate the demands of broader, longer, and more complex sensing environments. Designed to run unobtrusively in the phone’s background, it passively gathers data without requiring user interaction, an essential design choice that facilitates large-scale and low-maintenance longitudinal studies. To optimize battery life and improve inference quality, the system doesn’t collect raw accelerometer data. Instead, it uses the iOS Core Motion API and Google Activity Recognition API to categorize user activities (like being stationary, walking, running, cycling, or driving). Additionally, the app records detailed phone usage and location data, including phone lock/unlock events and periodic GPS coordinates, storing this information locally on the device. When an internet connection is available, the app securely transmits the stored data to our servers and then deletes it from the local storage.

Our backend system is composed of several elements: a HTTP server dedicated to participant enrollment, which facilitates the introduction of study purposes, online participant screening, user ID and login credential generation, and app distribution; a HTTP app portal service that connects with the sensing app and receives uploaded data; and separate data storage servers—one MySQL server for identifiable user demographic information and one MongoDB server for the anonymized sensing data, as typically mandated by IRB protocols. The data processing service extracts features from the raw data stored on the MongoDB server, computing behavioral attributes for downstream analysis. These attributes are also fed into an administrative dashboard, accessible via HTTP, where researchers can oversee study adherence and monitor the
1.2 Methodology: Smartphone Sensing Systems and More

health of the app.

1.2.2 Key Updates on the StudentLife Sensing Systems

We revamp the StudentLife sensing app to enhance its sensing efficiency and dependability, enabling it to support multi-year-long studies. Specifically, we adopt new app design, integration, and deployment strategies.

We build the iOS version of the StudentLife sensing app based on Voice over Internet Protocol (VoIP) push notifications. Continuous sensing on iOS has historically been challenging because of the restrictive nature of the environment. Because iOS does not support “real” multitasking, a third-party app will often have a limited amount of time available to execute when the system switches it to the background mode [12]. Existing mobile sensing applications rely on various background task declarations to resolve this. For example, the AWARE platform [18] claims to be a navigation program that constantly updates users on their location with the GPS sensor always on. Another iOS issue that can limit sensing apps is that missing data can arise if the background process is canceled (e.g., when insufficient memory is available), making the app unable to resume normally. We redesigned the iOS sensing system based on Voice over Internet Protocol (VoIP) push alerts to address these issues. By enabling our app to make phone calls over an internet connection instead of cellular service, we trigger the application’s sensing duty cycle with the help of oncoming traffic (i.e., push notifications to the app). Typically, our app is paused in the absence of such traffic. We send push notifications to the app from the backend, allowing it to wake up at regular intervals to run its sensing and inference pipeline. This technique has two benefits: first, the app is suspended and consumes
1.2 Methodology: Smartphone Sensing Systems and More

almost no energy when sensors are not used. Second, this technique considerably enhances the system’s robustness; even if an undesirable crash occurs, the app will resume functioning without user intervention.

We use the AppCenter [236] for the software development toolkit (SDK) management and app deployment. Specifically, the App Center is a continuous integration and delivery platform for iOS and Android phones, enabling fast, convenient and low burden app development and release cycles for pushing app updates to student phones. The SDK helps to monitor software crashes on the phone and allows researchers to push an update to users after fixing problems immediately. Next, we design the iOS version of the StudentLife sensing app based on Voice over Internet Protocol (VoIP) push notifications.

The backend system also includes a redesigned dashboard (as shown in Figure 1.1) used to quickly inspect and visualize the data collected from users, as well as port EMA and overall compliance rates. The dashboard is used daily by researchers responsible for “data sitting” the study to identify problems proactively. Figure 1.1 shows compliance data displayed in the dashboard. Figure 1.1a shows the last 30-days (only 12 days are displayed) of compliance, where each row corresponds to a participant. The page employs different colors to indicate different compliance rates (green indicates no missing data, yellow indicates some missing data, and red indicates no data), allowing us to quickly identify students in our study who are experiencing data collection issues and contact them to resolve the problem. If a certain UID (user identification) as shown in Figure 1.1a is “clicked”, the screen transitions to Figure 1.1b, which displays daily summaries of an individual student’s data quality. This includes the number of hours the phone is powered on, the number of hours of
1.2 Methodology: Smartphone Sensing Systems and More

collected activity labels, the number of hours of the GPS data available, the distance traveled, the number of EMA the user completed, and so forth.

![Table and figure captions](image)

(a) Last 30-day compliance visualization.

(b) Daily summary history of a particular user.

Figure 1.1: The StudentLife dashboard used to “datasit” the study.

1.2.3 Feature Extraction

We generate several features from the collected mobile sensing data. We list a few features which are typically computed by the backend system. The inclusion of spe-
1.2 Methodology: Smartphone Sensing Systems and More

cific features is contingent on the particular requirements of each study, leading to potential exclusions of some features and the addition of new ones. The nuances and adjustments made in feature selection across different studies will be elaborated upon in the subsequent chapters.

**Physical Activity.** Physical activity has been found to be associated with mental well-being. The MONARCA research [257] initially reported on data from mobile sensing and bipolar disorder, examining relationships between activity levels throughout the day and mental evaluation scores linked with the depressive spectrum. The findings in [2] suggest that circadian stability is a measure that can aid in the effective management of bipolar disorder. We use the iOS Core Motion API [164], and Google Activity Recognition API [125] to determine the physical activity of a student.

**Phone Usage.** Researchers [288] identified correlations between phone usage features from mobile phones and the severity of depressive symptoms, as measured by the 9-question Patient Health Questionnaire (PHQ-9) [187] in 40 subjects. PHQ-9 is a diagnostic tool used to screen for the presence and severity of depression. It rates depression based on the self-reported Patient Health Questionnaire. Previous research [85] indicates excessive smartphone usage is connected to depression or anxiety. In [163], researchers identified a positive relationship between smartphone screen time (e.g., phone unlock duration) and resting-state functional connectivity (RSFC) between brain regions associated with depression and antidepressant treatment response. The StudentLife app tracks the number of phone locks and unlocks students perform. We calculate both the total number of phone locks and unlocks, and the average time between phone locks and unlocks.
1.2 Methodology: Smartphone Sensing Systems and More

**Mobility.** Researchers discovered that mobility features are associated with depression. Wang et al. [347] found that mobility and location features act as proxy measures of decreasing interest or pleasure in activities. StudentLife sample GPS every 10 minutes to balance energy conservation and data quality. Raw GPS coordinates are first clustered using density-based spatial clustering with noise (DBSCAN) [101]. Following this, we calculate the number of unique locations and the distance traveled.

**Semantic Locations.** We can further assign semantics to places where participants visit. For example, in on-campus studies, we use a campus-wide map of buildings to categorize the semantics of locations, such as study areas, dorms, social spaces, gyms, Greek houses, and so on, based on their primary function to understand better how much time students spend at the various semantic locations across campus. Based on this semantic understanding of locations, we compute contextually aware behavioral features; for example, we can learn how long students use their phones (lock/unlock) while in study areas, dorms, etc.

**Sleep.** Sleep changes are one of the common symptoms associated with major depressive disorders [95]. Demirci et al. [20] discovered the association between sleep quality, depression and anxiety in 319 university students. We infer sleep duration, bedtime, and wake-up time using the method described in [63, 342], which had an accuracy of +/- 32 minutes to the ground truth.

**Ambient light.** (*From Android devices only*) The application measures the ambient light conditions of the user’s surrounding environment using light sensors on the phones. This can provide additional contextual information about the environment the user is in.
1.2 Methodology: Smartphone Sensing Systems and More

Phone calls and SMS. (*From Android devices only*) The application keeps track of SMS text message exchanges (both sent and received), as well as the number and length of phone calls. No written or audio material from these calls or texts was captured to protect privacy. We calculate the number of in/out SMS and calls, and the duration of in/out calls from the logs.

Conversation duration and frequency (*From Android devices and iOS devices before 2018: conversation data not available after 2018 due to iOS restrictions*) The study application uses the smartphone microphone to sample and gathers ambient sound in the device’s vicinity. A validated [362, 363, 361] in-situ privacy-preserving speech classifier determines when human conversations was nearby. To ensure privacy, no raw audio was captured on the device as part of the sensing system, but instead, the data was classified in the time and just the presence of voice was recorded. We compute the number of independent conversations and their duration.

1.2.4 Modeling Methods

Our studies employ a diverse array of modeling methods tailored to the specific analytical needs of each research question. To discern group differences, we utilize appropriate statistical tests. For scenarios involving multiple measurements within individuals, we employ models that account for within-individual dependencies, such as Generalized Linear Mixed Models (GLMM) [228] and Generalized Estimating Equations (GEE) [48].

For predictive analyses aiming to estimate the ground truth, our initial approach involves traditional machine learning techniques. These methods necessitate averaging each behavioral feature over the study period, given their inherent limitations.
1.2 Methodology: Smartphone Sensing Systems and More

in directly processing time-series data. Traditional machine learning models are preferred initially for their ability to yield more interpretable results, such as through the coefficients derived from linear models or feature importance scores from tree-based models.

When traditional machine learning methods fall short in achieving satisfactory outcomes, we pivot to deep learning techniques, such as LSTM and self-attentive networks. These advanced models are adept at learning complex representations that capture the temporal dynamics of sensor data, offering enhanced predictive capabilities. Although deep learning models typically lack inherent interpretability, we leverage model-agnostic interpretation tools to extract and understand the features and insights they generate.

1.2.5 Ground Truth

In our mobile sensing studies, the objective is to leverage smartphone data to predict specific outcomes, which, in our case, encompass personality traits, social functioning scales, auditory verbal hallucination (AVH) severity, and the mental well-being of college students. The methods used to collect ground truth data for these outcomes vary across studies. For instance, the Big Five Inventory (BFI) scores [172], a self-reported measure, are used to determine personality traits at the time of study enrollment. The Hamilton Program for Schizophrenia Voices Questionnaire (HPSVQ) [333] assesses the severity of AVH. Periodic assessments, such as the Social Functioning Scale (SFS) questionnaire [33] administered clinically every three months, are employed to monitor dynamic changes. Another important method involves in-situ mobile ecological momentary assessments (EMA) [305], integrated within the mobile
1.3 Problem Statement

sensing app, prompting users to complete short surveys at different times of the day. An example of this is the weekly self-reported Patient Health Questionnaire-4 (PHQ4) data collected via mobile EMA, providing timely insights into the participants’ mental health status.

1.3 Problem Statement

After discussing methodology in our smartphone sensing studies in general, we present the specific problems. We aim to expand the capabilities of mobile sensing to understand and interpret complex human behaviors and conditions in diverse and evolving environments. We seek to push the boundaries of current mobile sensing technologies, making them more adaptable, inclusive, and capable of capturing the nuanced dynamics of human life. Our investigation is structured around five principal areas. The problem statement and the major outcome of each area is as follows.

1.3.1 Broadening the Scope of Mobile Sensing for Personality Assessment

Personality traits describe individual differences in patterns of thinking, feeling, and behaving (“between-person” variability). But individuals also show changes in their own patterns over time (“within-person” variability). Existing approaches to measuring within-person variability typically rely on self-report methods that do not account for fine-grained behavior change patterns (e.g., hour-by-hour). We explore the potential of mobile sensing to assess personality traits, aiming to understand how daily behaviors captured through smartphones can reflect underlying personality struc-
1.3 Problem Statement

Specifically, we use passive sensing data from mobile phones to examine the extent to which within-person variability in behavioral patterns can predict self-reported personality traits. Data were collected from 646 college students who participated in a self-tracking assignment for 14 days. To measure variability in behavior, we focused on 5 sensed behaviors (ambient audio amplitude, exposure to human voice, physical activity, phone usage, and location data) and computed 4 within-person variability features (simple standard deviation, circadian rhythm, regularity index, and flexible regularity index). We identified a number of significant correlations between the within-person variability features and the self-reported personality traits. Finally, we designed a model to predict the personality traits from the within-person variability features. Our results show that we can predict personality traits with good accuracy. The resulting predictions correlate with self-reported personality traits in the range of $r = 0.32, \text{MAE} = 0.45$ (for Openness in iOS users) to $r = 0.69, \text{MAE} = 0.55$ (for Extraversion in Android users). Our results suggest that within-person variability features from smartphone data has potential for passive personality assessment.

1.3.2 Utilizing Mobile Sensing for Social Functioning Analysis

Impaired social functioning is a symptom of mental illness (e.g., depression, schizophrenia) and a wide range of other conditions (e.g., cognitive decline in the elderly, dementia). Today, assessing social functioning relies on subjective evaluations and self assessments. We examine the capacity of mobile sensing data to provide insights into individuals’ social functioning, particularly among populations with mental health disorders, and how these interactions change in different contexts.
1.3 Problem Statement

Specifically, we propose a different approach and collect detailed social functioning measures and objective mobile sensing data from N=55 outpatients living with schizophrenia to study new methods of passively accessing social functioning. We identify a number of behavioral patterns from sensing data, and discuss important correlations between social function sub-scales and mobile sensing features. We show we can accurately predict the social functioning of outpatients in our study including the following sub-scales: prosocial activities (MAE = 7.79, r = 0.53), which indicates engagement in common social activities; interpersonal behavior (MAE = 3.39, r = 0.57), which represents the number of friends and quality of communications; and employment/occupation (MAE = 2.17, r = 0.62), which relates to engagement in productive employment or a structured program of daily activity. Our work on automatically inferring social functioning opens the way to new forms of assessment and intervention across a number of areas including mental health and aging in place.

1.3.3 Enhancing Auditory Verbal Hallucination (AVH) Severity Prediction through Extended Sensing Signals

Mobile phone sensing is increasingly being used in clinical research studies to assess a variety of mental health conditions (e.g., depression, psychosis). However, in-the-wild speech analysis – beyond conversation detecting – is a missing component of these mobile sensing platforms and studies. We investigate the integration of additional sensory data, such as text and audio from voice diaries, with mobile sensing to improve the prediction and understanding of AVH severity.

Specifically, we augment an existing mobile sensing platform with a daily voice diary to assess and predict the severity of auditory verbal hallucinations (i.e., hear-
1.3 Problem Statement

Ining sounds or voices in the absence of any speaker), a condition that affects people with and without psychiatric or neurological diagnoses. We collect 4809 audio diaries from N=384 subjects over a one-month-long study period. We investigate the performance of various deep-learning architectures using different combinations of sensor behavioral streams (e.g., voice, sleep, mobility, phone usage, etc.) and show the discriminative power of solely using audio recordings of speech as well as automatically generated transcripts of the recordings; specifically, our deep learning model achieves a weighted f-1 score of 0.78 solely from daily voice diaries. Our results surprisingly indicate that a simple periodic voice diary combined with deep learning is sufficient enough of a signal to assess complex psychiatric symptoms (e.g., auditory verbal hallucinations) collected from people in the wild as they go about their daily lives.

1.3.4 Predicting Mental Well-being of College Students with a Focus on First-Generation Students

The transition from high school to college is a taxing time for young adults. New students arriving on campus navigate a myriad of challenges centered around adapting to new living situations, financial needs, academic pressures and social demands. First-year students need to gain new skills and strategies to cope with these new demands in order to make good decisions, ease their transition to independent living and ultimately succeed. In general, first-generation students are less prepared when they enter college in comparison to non-first-generation students. This presents additional challenges for first-generation students to overcome and be successful during their college years. This problem area focuses on leveraging mobile sensing to predict the mental well-being of college students, with a special emphasis on identifying and
1.3 Problem Statement

understanding the unique behavioral factors of first-generation college students.

Specifically, we study first-year students through the lens of mobile phone sensing across their first year at college, including all academic terms and breaks. We collect longitudinal mobile sensing data for \( N = 180 \) first-year college students, where 27 of the students are first-generation, representing 15\% of the study cohort and representative of the number of first-generation students admitted each year at the study institution, Dartmouth College. We discuss risk factors, behavioral patterns and mental health of first-generation and non-first-generation students. We propose a deep learning model that accurately predicts the mental health of first-generation students by taking into account important distinguishing behavioral factors of first-generation students. Our study, which uses the StudentLife app, offers data-informed insights that could be used to identify struggling students and provide new forms of phone-based interventions with the goal of keeping students on track.

1.3.5 Addressing Domain Drift and Model Degradation through Adaptation Technologies

College students have high prevalence of depression. Behind the success of machine learning models for depression diagnosis and detection using mobile sensing, the community yet does not know how long such models can last (namely, how fast they degrade). We study the effects of domain drift over the college life of students and explore the application of adaptation technologies to mitigate model degradation, ensuring that mobile sensing frameworks remain effective over long periods.

Specifically, we collected passive sensing data and self-reported surveys from 215 undergraduate student at Dartmouth College, spanning from September 2017 to June
1.4 Protection of Human Subjects

2022. We detect the drifts in students’ behavior and concept and analyze the degradation in model performance. We show that applying domain adaptation strategies can make depression detection tasks more robust to changes in the dataset over time.

1.4 Protection of Human Subjects

The personality study has been approved by the Office of Research Support and Compliance at The University of Texas at Austin. The social functioning study has been approved by the Committees for the Protection of Human Subjects at Dartmouth College and Human Services and the Institutional Review Board at Zucker Hillside Hospital. The AVH study has been approved by the Institutional Review Board of the of the University of Washington and Dartmouth College. The first-generation and domain drift study has been approved by the Institutional Review Board of Dartmouth College.

In all studies, we implemented key design features to safeguard participants’ privacy, including: (a) obtaining consent for app installation and data tracking, (b) allowing opt-out anytime, (c) using random identifiers for data, (d) anonymizing data, (e) leveraging app permissions, (f) securing data transfer with SSL encryption, (g) protecting servers with password logins, (h) storing physical servers in locked spaces accessible only during business hours, (i) separating personally identifiable information from sensing data across different servers, and (j) restricting data access to internal researchers, aligned with IRB guidelines and participant consent.

The personality study is supported by the National Science Foundation (NSF) Award BCS-1520288. The social functioning study is supported by the National Institute of Mental Health, grant number R01MH103148. The AVH study is supported by
1.4 Protection of Human Subjects

the National Institute of Mental Health (NIMH), grant number R01MH112641. The first-generation and domain drift study are supported by the National Institute of Mental Health (NIMH), grant number 5R01MH059282.

The thesis is structured as follows: Chapter 2 discusses mobile sensing for personality assessment, while Chapter 3 focuses on its use in analyzing social functioning. Chapter 4 explores enhancing AVH severity prediction with extended sensing signals. Chapter 5 covers the prediction of college students’ mental well-being, emphasizing first-generation students. Chapter 6 tackles domain drift and model degradation, proposing adaptation technologies. Each chapter addresses a distinct area of our comprehensive mobile sensing framework in diverse environments.
Chapter 2

Sensing Behavioral Change over Time: Using Within-Person Variability Features from Mobile Sensing to Predict Personality Traits

2.1 Introduction

Personality psychology focuses on examining individual differences in people’s thoughts, feelings, and behaviors. Compared to the amount of research on people’s thoughts and feelings, considerably less research has examined how people behave in the context of everyday life (e.g., daily levels of physical activity, sociability, places visited). Traditionally, research examining individual differences has focused on between-person variability in mean levels of such behaviors. For example, people who are more ex-
troverted can be characterized by higher mean levels of talkativeness, compared with people who are less extroverted. However, people also vary in the extent to which their own behavioral patterns change over time, which is referred to as *within-person* variability. For example, a person who is described as being more extroverted, may show great variability (e.g., socializing little during the week, but a great amount during the weekend) or stability (e.g., socializing a similar amount every day of the week) in their behavior patterns when considered over time.

Past research has shown within-person variability to be linked to various psychological characteristics, such as a person’s affective states [178], mental well-being [225] and personality trait ratings [110]. However, past research often relied on a person’s capacity to accurately recall their daily experiences retrospectively [242, 241], a task that is challenging and time-consuming. Individuals may also intentionally under-report or over-estimate some of their behaviors [328]. Fortunately, smartphone sensing methods [193] are set to overcome these barriers by unobtrusively measuring behavioral patterns continuously over time and thereby allowing us to understand the fine-grained within-person variability in behavioral patterns.

In this chapter, we present a new approach to capture within-person variability in behaviors using mobile sensing with the goal of assessing and predicting self-reported personality traits. We use the Big Five model [172], which describes 5 major personality trait dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Openness is a personality trait that describes the extent to which a person is imaginative and insightful. Conscientiousness is a trait that describes the extent to which a person is thoughtful, shows impulse control, and engages in goal-directed behaviors. Extraversion is a trait that describes the extent to which a person
2.1 Introduction

is excitable, social, talkative, and exhibits emotional expressiveness. Agreeableness is a trait that describes the extent to which a person is trusting, altruistic, kind, and engages in prosocial behaviors. Neuroticism is a trait that describes the extent to which a person is moody and emotionally unstable.

To examine how within-person variability in behavioral patterns are related to trait ratings, we consider the following everyday behaviors inferred using mobile sensing [141]: social interactions (i.e., how much a person socializes), physical activity and mobility (i.e., how physically active they are, and how many different places they spend time in), daily activities (i.e., how often they use their phone), and situational information (i.e., noisiness of their environment). To quantify within-person variability in lifestyle behaviors, we compute the following within-person variability metrics for each of the four everyday behaviors: standard deviation, circadian rhythm, regularity index, and flexible regularity index. We examine the connections between within-person variability metrics and personality traits; specifically, we pose the following broad research question: To what extent do day-to-day behavioral patterns of stability and change reveal a person’s personality traits? To answer this research question, we collected self-reported personality trait ratings as our measure of ground truth, along with the following sensing data as our measures of lifestyle behaviors: ambient audio amplitude levels (which indicates how quiet or noisy the environment is), exposure to human voice (relating to how social the user is), physical activity, phone usage and location data from 646 students using their Android and iPhones at the University of Texas at Austin (UTA) over a course of up to 2 weeks. The contributions of the chapter are as follows:

- We demonstrate for the first time how within-person variability patterns col-
lected passively by a comprehensive cross-platform (i.e., Android and iOS) mobile sensing app can be used to predict personality traits. Specifically, we measure and assess everyday behaviors including social interactions, physical movement, daily activity and situational information. Furthermore, we propose using the following measures of within-person variability: standard deviation, circadian rhythm, regularity index, and flexible regularity index. These measures capture behavioral variability from different perspectives (these measures are explained in more detail in Section 3).

- We identify a number of important associations between the within-person variability features and self-reported personality traits. Furthermore, we predict personality traits solely based on within-person behavior change features. Our results show that our proposed personality prediction model based on within-person variability features provides good estimation of personality traits, particularly for extraversion and agreeableness. For example, for Android users the leave-one-out model for predicting the extraversion trait achieves 0.55 of MAE, which is 0.24 (30%) lower than the average baseline and 0.5 (48%) lower than the random baseline; for iOS users the leave-one-out model for predicting the extraversion trait achieves 0.61 of MAE, which is 0.11 (15%) lower than the average baseline and 0.39 (39%) lower than the random baseline.

To the best of our knowledge, we are the first to explore how within-person variability patterns can be used to predict personality traits using features derived from mobile sensing. Our results pave the way for future research on this psychological topic. The structure of the chapter is as follows. First, we present related work on personality and mobile sensing research in Section 2.2, followed by a detailed de-
2.2 Related Work

In recent years, mobile sensing has demonstrated its potential as a tool for tracking and modeling human behavior [142, 144, 234, 288]. Equipped with unobtrusive sensors, smartphones can collect continuous sensing data that reveal individuals’ behavioral patterns and psychological states over long periods of time. For example, several studies have used smartphone sensing to continuously assess people’s mental health [17, 193, 273]. The StudentLife study [342] investigated the relationship between many types of smartphone data (e.g., conversation, sleep, activity, and co-location) and mental health outcomes (e.g., depression, stress, loneliness, and flourishing) in Dartmouth students during an academic term. Using the same dataset, Harari et al. [139] analyzed the changes of students’ activity and sociability behaviors over a term via the accelerometer and microphone sensors. Ben-Zeev et al. [24] used mobile phones to collect passive sensing data from smartphones and find schizophrenia relapse signals in location, activity, and exposure to conversation prior to patients experiencing relapses. Saeb et al. [288] reported depressive symptom severity correlates with mobility patterns and phone usage derived from smartphone data. They
2.2 Related Work

replicated their findings using the StudentLife [342] dataset [287]. Canzian and Musolesi [53] proposed a location routine index computed from smartphone location data, which was predictive depression severity. Abdullah et al. [2] reported using location features computed from smartphone data, such as distance traveled, conversation frequency, and non-stationary duration, to infer the social rhythm metric (SRM) [242] score, a widely used lifestyle regularity metric.

Several other studies have focused on inferring stability or variability in lifestyle behavioral patterns inferred from smartphone sensing data. For example, Abdullah et al. [3] computed daily rhythms related to sleep to measure well-being. Saeb et al [288] explored using circadian rhythm inferred from smartphone location data to assess depressive symptom severity. Ghandeharioun et al. [122] computed a sleep regularity index (SRI) using accelerometer data and show that SRI differs significantly between days with good and poor mental health. Mehtrotra and Musolesi [234] proposed a movement digital biomarker for monitoring emotional state, which measures the similarity between the sequences of visited places in a day.

There has also been an increasing amount of research focused on predicting personality traits from digital media data. Researchers have sought to predict personality using social network structures and interactions, showing that the behavioral data collected from social media platforms such as Facebook [372, 335, 211, 260] or Twitter [124, 123, 272] can be informative in the prediction of people’s personalities. Youyou et al. [372] showed that using Facebook Likes, a computer can predict participants’ personality more accurately than their Facebook friends. Moreover, the computer-made personality judgments had higher external validity when predicting life outcomes (e.g., substance use, political attitudes, and physical health).
et al. [260] built a predictive model of personality based on 66,732 Facebook users’ written language. The predicted personality scores correlate with the ground truth. Golbeck et al. [123] shows that Twitter users’ language use, sentiment, and Twitter use can be used to predict personality traits. However, these methods usually require access to extensive information about people’s online social networks.

Researchers have also utilized the information from mobile phones in personality assessment studies. Early studies tried to find relationships between personality traits and phone communications (i.e., phone calls and text messages) [49, 64, 83]. For instance, Montjoye et al. [82] used standard mobile phone logs (i.e., calls and texts) to predict users’ personality. Staiano et al. [315] collected call logs and Bluetooth proximity data from 53 subjects over 8 weeks to build a call network and a Bluetooth proximity network and used network characteristics to predict users’ personality traits. A recent study [340] explored using the StudentLife dataset, which comprises Wi-Fi, GPS, Bluetooth, accelerometer, and Piazza usage data to assess personality, finding significant correlations between behavior features Wi-Fi location based behavior features and personality traits. However, much of the aforementioned work used behavioral features focused on capturing ‘between-person’ variability in behavioral patterns. In this chapter, we focus on capturing and assessing within-person variability features from objective sensing behaviors, and show how these features can predict personality traits of smartphone users.

### 2.3 Within-person Variability Measurements

In what follows, we describe the passive smartphone sensing measures of behavior that we used to quantify behavioral variability. We first describe the behavioral data
2.3 Within-person Variability Measurements

collected (social interactions, movement and mobility, daily activity, and situational information), then we introduce the metrics that we used to quantify within-person variability (Standard Deviation, Circadian Rhythm, Regularity Index, Flexible Regularity Index).

2.3.1 Behaviors Inferred from Passive Smartphone Sensing

In what follows, we present the four aspects of daily behaviors that are captured through passive sensing using smartphones. We extend the StudentLife Android app [2, 194, 342, 346], which was originally used to capture students’ behaviors during a term and port it to the Apple iOS platform. The app measures daily social interactions, movement and mobility, daily activities and situational information by continuously collecting audio amplitude, ambient voice, participants’ physical activities, lock/unlock events and location coordinates.

Social interaction. We consider inferred ambient voice labels as a proxy for social interactions (i.e., being around conversation). We implement a conversation classifier on the phone to infer whether or not a 32ms audio frame is human voice. The voice classifier is implemented using a duty-cycled audio sensor that continuously runs on the smartphone. The voice classifier is the most energy consuming module in the app. To save energy, we set the duty cycle to be 1 minute on and 3 minutes off. It has been show in [342] that using this duty cycle we can achieve a balance between accuracy and resource usage. Further more, to preserve participant privacy, only the labels from the classifier are kept in the dataset and no raw human voices or speech content is recorded. Our classifier uses privacy preserving features [273, 342] and first determines if the frame contains speech and if so a higher level conversation
2.3 Within-person Variability Measurements

classifier determines if there are sufficient speech frames to indicate the start and later the end of a conversation. The frequency and duration of conversations are stored on the phone and uploaded for analysis. Note that the conversation classifier does not use speaker identification for privacy reasons and therefore only indicates if the user is in the presence of conversation rather than being an active speaker. Therefore, we consider our conversation inferences a proxy for social engagement. The speech/conversation classifier has been validated in several previous studies [273, 24, 342, 346].

Movement and mobility. The phone provides location data that allows us to understand users’ movement and mobility patterns. We find a user’s significant places (i.e., those places where users spent significant time) during the day and their associated dwell times (when the user arrives and leaves a location) by clustering the sampled coordinates during a day using density-based spatial clustering of applications with noise (DBSCAN) [101]. The DBSCAN algorithm groups the points that are close to each other and computes the center of the cluster. The center of the cluster is considered a significant location.

Daily activity. We consider two kinds of daily activity: physical activity and phone activity (i.e., phone usage). Personality traits are hypothesized to exert influence on physical activity through a health-behavior model [226, 279, 357]. Our app obtains activity inferences (i.e., stationary, walking, running, cycling, in vehicle) from the Android activity recognition API [125] and iOS Core Motion [164]. We compute the sedentary duration within every hour of the day using the phone’s physical activity inferences. Another aspect of daily activity we consider as providing signal is phone usage. An increasing number of researchers have shown that smartphone usage reflects
2.3 Within-person Variability Measurements

psychological well-being [35, 49, 179, 198, 213]. We compute the number of phone lock/unlock events and phone unlock duration to estimate the phone usage.

**Situational information.** We use ambient sound from the phone as a proxy for contextual information about the environment of users. Previous work [233] shows that participants’ personality traits are related to the quotidian manifestations derived from the sampled snippets of ambient sounds of users immediate environment. We periodically collect sound levels to measure the ambient sound environment. For privacy reasons we do not store any raw audio data. Rather, we compute the average sound amplitude over a period of one second so that the audio can not be reconstructed.

2.3.2 Within-Person Variability Metrics

Standard deviation is one of the simplest and most common approaches to measuring within-person variability. However, there may be other behavioural change metrics that better capture the patterns of change and stability in peoples’ everyday lives. These approaches could be more meaningful in better understanding within-person variability and its relationship to personality. To study this, we compute a number of variability measures including: the simple standard deviation, the circadian rhythm [3, 53, 288, 287], the regularity index and the flexible regularity index. We use these measures to assess the within-person variability of each behavior inferred using passive sensing. We first partition a day’s data into 24 one-hour periods and process the sensor data in an hourly fashion. For example, consider ambient audio amplitude, we compute the mean audio amplitude for each one-hour period; for voice, we compute the amount of conversation duration measured in one-hour periods; for
2.3 Within-person Variability Measurements

physical activity, we compute the sedentary duration across each hour period; for the phone usage, we compute the number of phone unlock events registered during each hour period. We do not preprocess location data. Next, we compute the four within-person variability metrics discussed above. Table 2.1 summarizes the within-person variability features used in this study. In what follows, we describe each metric in detail.

**Standard deviations [STD]** measure the variance in daily behaviors. We compute the STD over three epochs during a 24-hour period: day time (9am–6pm), evening (6pm-12am), and night (12am-9am) across all days of the week. Because people are likely to have different behavioral patterns during weekdays in comparison to weekends, we compute the STD for the three epochs (i.e. day, evening, night) for weekdays data only. We do not, however, compute the STD for weekends because of limited amounts of weekend data.

**Circadian rhythm [CR]** measures the strength with which a user follows a 24-hour rhythm in behaviors [287]. Humans have a biological clock that optimizes the physiology and behavior of organisms, hormonal secretion and mood [184]. However, people differ in their circadian rhythms. For example, previous studies have shown individual differences in the morning and evening related to personality [11, 276]. We compute the CR across the study period for hourly behavioral data using spectrum analysis. Specifically, we first use the least-squares spectral analysis [271] to transform the behavioral sensing data (i.e., physical activity, phone activity, ambient sound, ambient voice) from the time domain to the frequency domain. We then compute the ratio of energy that fall into the $24 \pm 0.5$ h period (which corresponds to $2\pi/(24 \pm 0.5) = (0.2565, 0.2674)$) over the total spectrum energy in the
2.3 Within-person Variability Measurements

24 ± 12 h period (which corresponds to $2\pi/(24 ± 12) = (0.1745, 0.5236)$) as the CR:

$$CR = \frac{\int_{2\pi/(24.5)}^{2\pi/(12.5)} psd(x)dx}{\int_{2\pi/(30)}^{2\pi/(12)} psd(x)dx}$$ (2.1)

where $psd(x)$ denotes the power spectral density at frequency bin $x$. The raw hourly signal is first aligned to zero mean before performing spectral analysis. Fig. 2.1 shows an example of the computed CR for ambient sound. The plots show the raw hourly ambient sound amplitude and the corresponding power spectrum for two participants selected from the study. The raw data show that participant (a) has a more pronounced 24-hour cycle for ambient sound than (b). As such, the CR value for participant (a) is 0.093 whereas the CR value for participant (b) is 0.040. For the mobility data, we use the steps described in [287] to compute the CR of mobility: we first generate the CR for the latitude and longitude values then combine them through $CR = \log(CR_{lat} + CR_{long})$.

**Regularity index [RI]** assesses the difference between the same hours across two different days. We first rescale the behavioral data for each participant to $[-0.5, 0.5]$, where -0.5 corresponds to the minimum value in the origin data and 0.5 corresponds to the maximum value. The product of two rescaled values is positive if the original values are close and negative if they are not similar. Subsequently, we define the regularity between day a and b as:

$$RI_{a,b} = \sum_{t-1}^{T} f(x_{i}^a) f(x_{i}^b)/T$$ (2.2)

where $S$ is the set of two-day pairs, $a, b$ are the two days in a two-day pair, $T = 24$ hours, and $x_{i}^a$ is the rescaled value in hour $t$ of day $a$. 

31
2.3 Within-person Variability Measurements

(a) Participant with high circadian rhythm value

(b) Participant with low circadian rhythm value

Figure 2.1: Circadian Rhythm analysis using audio amplitude sensing data from two participants.
2.3 Within-person Variability Measurements

We compute the mobility RI differently because unlike other behavioral data mobility data is nominal (i.e., a mobility data point represents a location). We therefore compare whether or not a user is at the same place in two days. The mobility RI is formally defined as:

\[
RI_{loc_{a,b}} = -\frac{1}{T_{loc}} \sum_{t=1}^{T_{loc}} g(c^a_t, c^b_t) / T_{loc}
\]  

(2.3)

where \( a, b \) represent two different days, \( t \) represent the time window in a day (a window lasts 10 minutes), \( T_{loc} \) is the available number of overlapped time windows in both days, \( c^a_t \) is the significant location id time \( t \) in the day \( a \), and \( g(m, n) \) indicate whether or not \( m = n \). A higher RI score indicates that the user visited similar places around a similar time of day in two given days.

We compute the average and range of the RI values from every possible pair within the following sets: (1) weekdays vs weekends, (2) within weekdays, (3) within all days.

Flexible regularity index [FRI] is an edit distance based (or Levenshtein distance [200]) measure to assess the difference between two days differently. An edit distance quantifies how dissimilar two strings are to one another by counting the minimum number of operations needed to transform one string to the other. Such operations include removing, inserting, or substituting one character in the string. Different operations may have different weights. The edit distance of the behavioral data in two days reveals how similar the behaviors in two days are. A lower edit distance (i.e., lower FRI) between different two days indicates more similar behaviors.

We compute the FRI as follows. First, we transform the behavioral sensing data into strings. Specifically, for behavioral data other than mobility, we label a one-hour chunk as ‘a’ if the mean sensor reading in this hour is within the bottom 25 percentile
2.3 Within-person Variability Measurements

of all data from this user; ‘c’ if the mean is within the top 25 percentile; and ‘b’ if
the mean is between the bottom 25 percentile and top 25 percentile. For mobility, we
use the significant location id to generate the mobility string in day. We define the
weights for each operation as shown in Table 2.2.

Table 2.1: Description of the features computed

<table>
<thead>
<tr>
<th>Feature category</th>
<th>Features computed</th>
<th>Description of feature</th>
<th>Sensing data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>std_night_all</td>
<td>STD on the data at night across all days</td>
<td>Ambient sound</td>
</tr>
<tr>
<td></td>
<td>std_day_all</td>
<td>STD on the data during day time across all days</td>
<td>Ambient voice</td>
</tr>
<tr>
<td></td>
<td>std_evening_all</td>
<td>STD on the data in the evening across all days</td>
<td>Physical activity</td>
</tr>
<tr>
<td></td>
<td>std_night_weekday</td>
<td>STD on the data at night across weekdays</td>
<td>Phone activity</td>
</tr>
<tr>
<td></td>
<td>std_day_weekday</td>
<td>STD on the data during day time across weekdays</td>
<td></td>
</tr>
<tr>
<td></td>
<td>std_evening_weekday</td>
<td>STD on the data at night across weekdays</td>
<td></td>
</tr>
<tr>
<td></td>
<td>circadian_all</td>
<td>Circadian rhythm from data across all days</td>
<td></td>
</tr>
<tr>
<td></td>
<td>circadian_weekday</td>
<td>Circadian rhythm from data across weekends</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

34
### 2.3 Within-person Variability Measurements

<table>
<thead>
<tr>
<th>Index (RI)</th>
<th>Measure Description</th>
<th>Data Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ri_weekday_vs_weekday_avg</strong></td>
<td>Measures the average hour-by-hour similarity between weekdays and weekends</td>
<td>Ambient sound</td>
</tr>
<tr>
<td><strong>ri_weekday_avg</strong></td>
<td>Measures the average hour-by-hour similarity across weekdays</td>
<td>Ambient voice</td>
</tr>
<tr>
<td><strong>ri_all_avg</strong></td>
<td>Measures the average hour-by-hour similarity across all days</td>
<td>Physical activity</td>
</tr>
<tr>
<td><strong>ri_weekday_vs_weekend_range</strong></td>
<td>Measures the range (i.e., the difference of the most similar pair and the most distinguished pair) of hour-by-hour similarity between weekdays and weekends</td>
<td>Phone activity</td>
</tr>
<tr>
<td><strong>ri_weekday_range</strong></td>
<td>Measures the range of hour-by-hour similarity between weekdays</td>
<td>Location</td>
</tr>
<tr>
<td><strong>ri_all_range</strong></td>
<td>Measures the range of hour-by-hour similarity across all days</td>
<td></td>
</tr>
</tbody>
</table>
2.4 Data Collection and Processing

We collected a dataset from 646 students at the University of Texas at Austin (UTA). The participants were enrolled in an online introductory psychology class across two semesters. As part of a course assignment, participants could self-track their lifestyle behaviors in exchange for personalized feedback using a tracking method of their choice: a mobile sensing app, email-based surveys, or a handwritten journal. Here we focus on the data collected from students who elected to use the mobile sensing app. Participants installed the data collection app on their phones and were asked
2.4 Data Collection and Processing

to participate for at least seven days. Participants were able to participate for up to fourteen days. Among the 646 participants, 117 used Android phones and 529 used iPhones. All participants complete the Big Five personality trait questionnaire [172] at the start of the study period, which serves as the ground truth for personality traits in this study.

Fig. 2.2 shows the overall system and study design. The complete system included the sensing app and cloud and was based on an earlier version of the StudentLife system [342]. We continuously collected behavioral passive sensing data from participants’ Android phones and iPhones. The data was then automatically uploaded to our secure server. The server processed the data, and generated personalized webpages of feedback reports, which were sent to students via email during the study. Those reports included personalized visualizations of the tracked behaviors as well as class average charts for comparison.

In what follows, we discuss how we processed the data in detail.

2.4.1 Data Inclusion Criteria

Data quality is crucial for analysis. Missing data across a day will adversely affect the accuracy of the within-person variability features. Therefore, we exclude days with less than 19 hours of sensing data. The 19-hour threshold is based on previous studies [342, 346], which balances the need for data quality and quantity. We also exclude participants who have less than 7 days of usable data (more than 19 hours of sensing data). 159 out of the 646 participants satisfy our data inclusion criteria and included in our analysis. Among them, 70 are Android users and 89 are iPhone users. The high data exclusion rate is mainly due to participants running the app for less
2.4 Data Collection and Processing

than 14 days. Other factors include the phone being powered down, turning location off, and stopping the app.

2.4.2 Big Five Personality Ground truth

We use the self-reported Big Five Inventory (BFI) [172] scores as our personality ground truth. The BFI measures the personality traits: openness, conscientiousness, extraversion, agreeableness and neuroticism. Fig. 2.3 shows the distributions of the Big Five scores of the included 159 participants. The distributions show that the values for all five traits approximate a normal distribution in this sample. Table 2.3 shows the mean and standard deviation of the five personality trait scores.

The mean of the trait scores are close to a score of 3 (the middle of the 1-5 range). The agreeableness score is close to 4, which is the highest mean of the trait scores,
2.4 Data Collection and Processing

Figure 2.3: Histograms of the Big Five scores. The X axis displays the value of the score, which ranges from 1 (lowest) to 5 (highest). The Y axis shows the number of individuals that fall into the specific score bins. The three rows show the distribution of the scores for all participants (first row), the Android (second row) and the iOS (third row) users.

Table 2.3: Distributions of the personality ground truth

<table>
<thead>
<tr>
<th>Big Five trait</th>
<th>Mean (std)</th>
<th>Android mean (std)</th>
<th>iOS mean (std)</th>
<th>t-test p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Openness</td>
<td>3.54 (0.62)</td>
<td>3.61 (0.61)</td>
<td>3.48 (0.62)</td>
<td>0.21</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>3.45 (0.65)</td>
<td>3.31 (0.63)</td>
<td>3.57 (0.65)</td>
<td>0.01</td>
</tr>
<tr>
<td>Extraversion</td>
<td>2.99 (0.90)</td>
<td>2.97 (0.92)</td>
<td>3.00 (0.88)</td>
<td>0.83</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>3.74 (0.63)</td>
<td>3.70 (0.63)</td>
<td>3.76 (0.63)</td>
<td>0.63</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>3.03 (0.79)</td>
<td>3.08 (0.79)</td>
<td>3.00 (0.80)</td>
<td>0.50</td>
</tr>
</tbody>
</table>

followed by openness, conscientiousness, neuroticism and extraversion. For most of the personality traits, minor differences are of small to negligible effect size. This is inline with a recent study [130] among a large multi-national (N = 1081) and a German-speaking sample (N = 2438). However, in our data Android users seem to be less conscientious than iOS users (t-test p = 0.01), which is in contrast to the findings...
in [130]. The reason could be the population and age: unlike a large multi-national user group with various occupations, our study participants are college students from one university. In addition, our population of users is smaller.

2.5 Assessing Personality using Within-person Variability Measures

In what follows, we present the results from association analysis and prediction of personality traits using passive sensing data from smartphones. We first extract all the within-person variability features and report linear correlations between the features and the self-reported personality scores. Then, we create a personality prediction model and analyze its prediction performance using only within-person variability features.

2.5.1 Association Analysis

We use the bivariate linear mixed model [228] to assess the relationship between the within-person variability features and Big Five personality traits. In our study, the sensing data come from two clusters: Android and iOS. There exist potential differences between two systems (e.g., both Google and Apple have their own physical activity classifiers and audio software development kits). Therefore, the sensed behavioral patterns are not independent. Linear mixed models are an extension of simple linear models to allow both fixed and random effects, and are particularly useful when there is non-independence in the data. Our association results are presented in Table 3.4. In order to address the multiple comparisons problem, we apply the
2.5 Assessing Personality using Within-person Variability Measures

Benjamini-Hochberg procedure (BH) [28, 30] to control the false discovery rate (FDR) in our exploratory regression analysis. The multiple comparisons problem arise when multiple simultaneous statistical tests are involved in the analysis, which may lead to erroneous discoveries. We present associations with $p < 0.05$ and mark associations that have FDR $< 0.1$ and FDR $< 0.05$.

In what follows, we discuss our results as they relate to the personality traits.

**Openness.** We find that four within-person variability features are positively associated with the openness trait. They are as follows: (1) the deviation in ambient sound on weekday (i.e. Monday-Friday) evenings (6pm-12am); (2) the deviation in physical activity in the evening across all days (i.e. Monday-Sunday); (3) the deviation in physical activity in the evening of weekdays; and finally (4) the range of the regularity index in ambient sound during weekdays. Our results indicate that participants who have various activities during the weekday evenings are more likely to be more open to new experiences. On several weekdays they stay in similar environments at the same hour of each day. However, on some other weekdays they spend time in very different environments (as captured by ambient sound).

We find that two within-person variability features are negatively associated with the openness trait. They are (1) the deviation in ambient voice between 9am - 6pm across all days; and (2) the deviation in ambient voice between 9am - 6pm on weekdays. Assuming ambient voice is a proxy for social interaction, our results indicate that people who have changing patterns in interaction with others during the daytime are likely to be less open to new experiences.

**Conscientiousness.** We find that four within-person variability features are positively associated with the conscientiousness trait. These are (1) the average of the
2.5 Assessing Personality using Within-person Variability Measures

flexible regularity index in ambient voice on weekdays; (2) the circadian rhythm in phone usage across all days; (3) the circadian rhythm in phone usage on weekdays; and (4) the flexible regularity index in physical activity. Our results indicate that participants who spend weekdays engaged in more regular social interactions (with hours slightly shifted), and who show more regular phone usage during the day, and have regular patterns of physical activity during weekdays are more likely to have good impulse control and goal-directed behaviors.

We find that five within-person variability features are negatively associated with the conscientiousness trait. These are (1) the deviation of exposure to ambient voice on weekday evenings; (2) the regularity index in locations between weekdays and weekends; (3) the regularity index in locations across all days; (4) the flexible regularity index in locations between weekdays and weekends; and finally (5) the flexible regularity index in locations across all days. This leads us to believe that individuals who have unstable social interactions during weekday evenings, and more overlap between weekdays and weekends in terms of their location routines are likely to be less conscientious.

Extraversion. We find that five within-person variability features are positively associated with the conscientiousness trait. These are (1) the circadian rhythm in ambient sound across all days; (2) the circadian rhythm in phone usage across all days; (3) the circadian rhythm in phone usage on weekdays; (4) the deviation in physical activity in the evening across all days; and finally (5) the deviation in physical activity in the evening on weekdays. This suggests that people who are more extraverted are, interestingly, more likely to follow a 24-hour rhythm with regards to the environments they spend time in (as captured by ambient sound) as well as their smartphone usage.
2.5 Assessing Personality using Within-person Variability Measures

These participants also tend to have various activities during the evening – similar to the patterns observed for openness.

In addition, we find that five within-person variability features are negatively associated with the conscientiousness trait. These are (1) the deviation in exposure to ambient voice during the daytime (9am-6pm) across all days; (2) the deviation in exposure to human voices during the evening (6pm-12am) across all days; (3) the deviation in exposure to human voices during the daytime (9am-6pm) on weekdays; (4) the deviation in exposure to human voices during the evening (6pm-12am) on weekdays; and finally (5) the deviation of phone usage in the evening periods across all days. These results tell us that more extroverted individuals are less likely to socialize in a changing pattern. Rather, they are more likely to maintain a stable pattern of interpersonal communication over the week, particularly in the evenings. Similarly, this also applies to their phone usage during evening periods.

**Agreeableness.** We find a large number of within-person variability features are positively associated with the agreeableness trait. These are (1) the range of the flexible regularity index in ambient sound; (2) the range of the flexible regularity index in phone usage on weekdays; (3) the circadian rhythm in physical activity; (4) the flexible regular index in physical activity on weekdays; (5) the flexible regularity index in physical activity across all days; (6) the regularity index in physical activity on weekdays; (7) the range of regularity index; and finally (8) the flexible regularity index in locations on weekdays. These results indicate more agreeable participants are more likely to have to follow a 24-hour rhythm in physical activity. They also tend to have more regular physical activity patterns – based on hour-by-hour comparisons between days. They are more likely to change their ambient sound environment and
2.5 Assessing Personality using Within-person Variability Measures

location routines for some weekdays, while they keep them unchanged on other days.

In addition, we find a number of within-person variability features that are negatively associated with the agreeableness trait. These are (1) the deviation in phone usage during the night (12am-9am) on weekdays; (2) the deviation in physical activity during the night across all days; (3) the deviation in physical activity during the night period on weekdays; (4) the range of the flexible regularity index in physical activity on weekdays; and finally (5) the range of the regularity index in physical activity on weekdays within weekdays, between weekdays and weekends, and across all days. This seems to tell us that people who are agreeable show less deviation in the time they use their phone during the night. They maintain high regularity in physical activity during the study, as indicated by the smaller range and higher average of their measured regular index.

Neuroticism. Neuroticism indicates moodiness and emotional instability. In our study, we do not see any within person variability features significantly associated with the neuroticism trait.

To sum up, the association analysis supports our hypothesis that within-person behavior change patterns derived directly from smartphone sensing data are related to self-reported personality traits. Four of the five personality traits were associated with different types of within-person variability features. However, some features associated with multiple personality traits. For example, the higher deviation in physical activity during the evening period was associated with being more open to new experiences and being more extroverted; the higher regularity in physical activity on weekdays was associated with being more conscientious and more agreeable.
2.5 Assessing Personality using Within-person Variability Measures

Table 2.4: Big Five traits in related to regularity using mixed-effect model.

<table>
<thead>
<tr>
<th>Big Five trait</th>
<th>Association</th>
<th>Related within-person regularity features (N=159)</th>
</tr>
</thead>
<tbody>
<tr>
<td>openness</td>
<td>(+)</td>
<td>sound_std_evening_weekday,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>stationary_std_evening_all,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>stationary_std_evening_weekday,</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>sound_ri_weekday_range</strong></td>
</tr>
<tr>
<td></td>
<td>(-)</td>
<td>voice_std_day_all, voice_std_day_weekday</td>
</tr>
<tr>
<td>conscientiousness</td>
<td>(+)</td>
<td>voice_fri_weekday_avg, lock_circadian_all,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lock_circadian_weekday,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>stationary_fri_weekday_avg</td>
</tr>
<tr>
<td></td>
<td>(-)</td>
<td>voice_std_evening_weekday,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>location_ri_weekday_vs_weekend_avg,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>location_ri_all_avg,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>location_fri_weekday_vs_weekend_avg,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>location_fri_all_avg</td>
</tr>
<tr>
<td>extraversion</td>
<td>(+)</td>
<td>sound_circadian_all, lock_circadian_all*,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lock_circadian_weekday,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>stationary_std_evening_all,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>stationary_std_evening_weekday</td>
</tr>
<tr>
<td></td>
<td>(-)</td>
<td>voice_std_day_all, voice_std_evening_all**,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>voice_std_day_weekday,</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>voice_std_evening_weekday</strong>,**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lock_std_evening_all</td>
</tr>
</tbody>
</table>
### 2.5 Assessing Personality using Within-person Variability Measures

<table>
<thead>
<tr>
<th>agreeableness</th>
<th>(+) sound_fri_all_range, lock_fri_weekday_range, stationary_circadian_all, stationary_fri_weekday_avg, stationary_fri_all_avg, stationary_ri_weekday_avg, location_ri_weekday_range, location_fri_weekday_range</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(-) lock_std_night_weekday, stationary_std_night_all, stationary_std_night_weekday, stationary_fri_weekday_range, stationary_ri_weekday_vs_weekend_range, stationary_ri_weekday_range, stationary_ri_all_range</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>neuroticism</th>
<th>(+)</th>
<th>(-)</th>
</tr>
</thead>
</table>

\[ p < 0.05; \textbf{bold} p < 0.01, ^*FDR < 0.1, ^{**FDR} < 0.05 \]

#### 2.5.2 Prediction Analysis

We use Gradient Boosted Regression Trees (GBRT) [115, 262] to predict the self-reported Big Five personality scores. GBRT is an ensemble method that trains and combines several weak regression trees to make more accurate and robust predictions. It builds base estimators (i.e., regression trees) sequentially. Each estimator tries to reduce the bias of the previously combined estimators. By doing so, in each stage a new regression tree is trained on the negative gradient of the loss function. GBRT is
less sensitive to outliers and robust to overfitting [99]. Another advantage inherited from the tree based model is that it computes feature importance measures, which can be used for feature selection.

We have a total of 96 features (see Table 2.1) and a relatively small number of training examples (70 Android and 89 iOS). We reduce the feature space dimensionality using the importance vector generated from GBRT. GBRT computes feature importance by averaging the number of times a particular feature is used for splitting a branch across the ensemble trees. Higher values correspond to higher importance. We select features with a feature importance value higher than the mean importance iteratively. We repeat this process until we get no more than 9 features. Our heuristic of selecting 9 features is based on experiments in which we find we get higher training errors with a lower or higher threshold.

We train and test models separately among the Android and iOS users. We take this strategy on the basis of our observation that a model trained from the mixed dataset gives poorer predictions, even if we normalize the features separately among Android and iOS users. We believe this is due to the differences (i.e., heterogeneities) between Android and iOS devices, i.e., the accuracy of sensed behaviors may be influenced by the sensors and algorithms on different platforms. For example, activities are derived from the Google Activity Recognition API [125] on Android phones and from the iOS Core motion API [164] on iPhones; the values of sound amplitude in the same environment may be distinct between the two platforms; the conversation classifier may have potentially small differences because of the different platforms (e.g., different microphones and audio APIs). While we have designed and implemented our sensing algorithms on both platforms to take known differences between the iOS
2.5 Assessing Personality using Within-person Variability Measures

and Android platforms into account there are still no widely accepted techniques for equalizing sensing data from these different platforms. Scaling the values to standard normal distributions separately on both platforms does not result in an ideal solution - it violates the observation that Android and iOS users may have different behaviors and phone usage patterns [152, 222, 27, 332, 210]. Adding a binary feature DeviceType to control for the platform of the device, does not result in better performance when there are no significant differences for most of the personality traits (Table 2.3). Specifically, in GBRT the tree grows greedily in a top-down fashion using binary splits. For each tree node, the split minimizing the objective is chosen. The DeviceType may not be selected at the root of regression trees due to the similar means of ground truth; i.e., the personality trait scores in the two groups. Taking these challenges into account we opted to divide the data into Android and iOS users. We train different models on these two groups.

Usually, with a small sample set leave-one-out cross validation [149] would be the best option to show the performance of the personality trait prediction model. This technique is widely used in estimating the performance of the model from a small human-centered dataset in existing studies [343, 170, 14, 375]. Leave-one-out has low bias, because each fold uses almost the entire dataset as a training set [192]. However, as a result, the estimation is also very specific for this particular dataset. This could results in high variance compared to the same model’s performance on new datasets. Therefore, we also use five-fold cross validation. For each personality trait, we use ten times five-fold cross validation and report the average.

We compare our predictive models with two baseline models. The baseline 1 model takes the average of the scores as the predicted value. This is the most basic regression
with only an intercept. The baseline 2 model randomly generates a sample from the already known distribution (i.e., the distribution of the personality trait scores) and uses it as the predicted value. We validate our prediction model using the Mean Absolute Error (MAE), the root mean squared error (RMSE), the Pearson correlation and the R-squared value. MAE and RMSE describe the bias of the predictions; the Pearson’s r describes how well the predictions are associated with the ground truth; and the R-squared value measures the goodness of fit by indicating how much of the variance our model explains [51].

Table 2.5 shows the performance of the prediction model purely based on within-person variability features. The models perform better than the two baselines, and capture considerable variance of the original distribution. The predicted personality score is highly correlated with the ground truth. Our model works better in predicting the extraversion and agreeableness traits. For example, for Android users the leave-one-out model for predicting the extraversion trait achieves 0.55 of MAE, which is 0.24 (30%) lower than the average baseline model and 0.5 (48%) lower than the random baseline model. For iOS users the leave-one-out model for predicting the extraversion trait achieves 0.61 of MAE, which is 0.11 (15%) lower than the average baseline model and 0.39 (39%) lower than the random baseline model. Our predictive model is less effective at predicting neuroticism, which is in line with the fact that we did not find features associated with neuroticism, as discussed in Section 5.1.
2.5 Assessing Personality using Within-person Variability Measures

<table>
<thead>
<tr>
<th>System</th>
<th>Big Five trait</th>
<th>Baseline 1 MAE/RMSE</th>
<th>Leave one out MAE/RMSE</th>
<th>5-fold cross validation MAE/RMSE</th>
<th>r²</th>
<th>corr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Openness</td>
<td>0.47/0.60</td>
<td>0.68/0.85</td>
<td>0.40/0.51</td>
<td>0.54</td>
<td>0.287</td>
</tr>
<tr>
<td>Android</td>
<td>Conscientiousness</td>
<td>0.48/0.62</td>
<td>0.69/0.88</td>
<td>0.41/0.54</td>
<td>0.53</td>
<td>0.252</td>
</tr>
<tr>
<td></td>
<td>Extraversion</td>
<td>0.79/0.92</td>
<td>1.05/1.28</td>
<td>0.49/0.62</td>
<td>0.46</td>
<td>0.366</td>
</tr>
<tr>
<td></td>
<td>Agreeableness</td>
<td>0.49/0.62</td>
<td>0.70/0.88</td>
<td>0.40/0.49</td>
<td>0.53</td>
<td>0.223</td>
</tr>
<tr>
<td></td>
<td>Neuroticism</td>
<td>0.66/0.79</td>
<td>0.90/1.11</td>
<td>0.46/0.203</td>
<td>0.20</td>
<td>0.117</td>
</tr>
<tr>
<td></td>
<td>Openness</td>
<td>0.50/0.62</td>
<td>0.69/0.87</td>
<td>0.45/0.59</td>
<td>0.46</td>
<td>0.072</td>
</tr>
<tr>
<td>iOS</td>
<td>Conscientiousness</td>
<td>0.50/0.65</td>
<td>0.72/0.91</td>
<td>0.46/0.59</td>
<td>0.42</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>Extraversion</td>
<td>0.72/0.88</td>
<td>1.00/1.24</td>
<td>0.48/0.63</td>
<td>0.40</td>
<td>0.143</td>
</tr>
<tr>
<td></td>
<td>Agreeableness</td>
<td>0.69/0.88</td>
<td>0.69/0.88</td>
<td>0.40/0.51</td>
<td>0.57</td>
<td>0.143</td>
</tr>
<tr>
<td></td>
<td>Neuroticism</td>
<td>0.64/0.79</td>
<td>0.90/1.12</td>
<td>0.40/0.158</td>
<td>0.20</td>
<td>0.100</td>
</tr>
</tbody>
</table>
Fig. 2.4 illustrates the predicted personality trait score (y-axis) and the self-reported ground truth (x-axis) for each participant using the leave-one-out method. The blue line indicates the ideal model where the predicted value is equal to the ground truth. The red points above the blue line are overestimated and the points below it are underestimated. Since our training set is normally distributed, the model gets reinforced in the center area and has bigger absolute errors on the two sides. Even though, we see that the model can still capture the trend and variance of the ground truth distribution. Because traits measured using the Big Five Inventory are designed to be treated as continuous variables, we do not conduct binary classifications. However, the plot indicates that if we did perform a binary classification to distinguish people with higher or lower scores on some traits, we would also achieve good performance results, particularly when predicting the extraversion and agreeableness traits.

2.6 Limitations

The current study has a number of limitations that need to be addressed in future research. First, further research is needed to show the data collected from Android and iOS devices can be correctly merged. In our prediction model we take a conservative approach and separate out the Android and iOS groups. We find a model trained from the mixed dataset gives worse prediction, even if we normalize the features separately among Android and iOS users. This could be because of the accuracy of sensed behaviors are influenced by the sensors and algorithms on different platforms, as discussed earlier in the chapter. If so, work is needed to mitigate the impairments caused by features collected on different platforms. This is an interesting and important area
2.6 Limitations

Figure 2.4: Predicted values of personality traits and ground truth.

of research in passive sensing on different devices and benchmarking norms between devices. Another possible explanation is that there are real differences between users who select different platforms; that is, there could exist different baseline personality
2.6 Limitations

trait expressions across the different user groups - which might in turn be explained by advertising strategies, pricing and the brand personalities of the companies behind the respective operating systems and major phone manufacturers using them. However, further work is needed to explore to what extent more fundamental differences in the operating systems, and how they get used by and interact with the phone users influence these differences, as well as whether the same disparities are observed across samples in other, non-student populations.

Second, further work is needed to explore which sensor-based features are the strongest predictors of personality traits. There might be some other powerful features which can better represent the changes in lifestyle that is more predictable in estimating personality other than the metrics we have used: that is, standard deviation, circadian rhythm, and two regularity indexes. As sensing devices become more powerful and additional sensors become available, the research on passive personality assessment may identify other variability features that are powerful predictors of behavior and personality.

Third, it is unclear how sensing restrictions on the iOS platform influenced our sample. The original sample consisted of 646 students, out of which 117 (18%) were Android users and 529 (82%) reported being iOS users. However, only about a quarter – 159 participants – could be included in the analyses presented in this chapter – that is participants that had more than 7 days of over 19 hours of sensing data. This is a conservative inclusion criteria we used in prior studies. Of those, 70 (44%) were Android users and 89 (56%) were iOS users. It is unclear to what extent the comparatively small subset of iOS users with sufficient amounts of sensor data compare to their excluded peers. Typically, Android is more open to continuous passive sensing
2.6 Limitations

even though new versions of the Android OS are placing more restrictions. Apple’s iOS has been more closed to continuous passive sensing by imposing restrictions on sampling rates and access to sensing data. As a result the more restrictive iOS environment limits data gathering. In our study, some iOS users reported the app was inadvertently terminated requiring us to troubleshoot the issues throughout the data collection process. Despite these challenges, we are confident that the actions we took to mitigate problems (e.g., conservative inclusion criteria, matching data quality collected across platforms, balance of iOS and Android users) has lead to a good quality dataset.

Fourth, further work is required to balance the accuracy and resources usage. We collected an extensive amount of data via passive sensing for capturing as various and accurate behaviors as possible. According to our survey [144] after this study, participants were satisfied with the self-tracking assignment using mobile phones. The average levels of satisfaction were 3.70 (Android) and 3.92 (iOS), respectively (from 1 Very unsatisfied to 5 Excellent). Besides, 61% participants reported they did not feel uncomfortable using the app at all. However, 53% Android and 28% iOS users noticed the draining phone batteries, which indicated the biggest obstacle of allowing the adoption in real life scenarios. As a follow-up, we tested the power consumption by turning on and off each sensing component (i.e, activity detection, accelerometer, in-situ voice classifier, GPS location and scheduled data uploading) on factory-reset Android and iOS phones. Our tests show that the voice classifier and collecting raw accelerometer data are the major causes of energy cost. We decide to decrease the resource usage by stopping collecting the accelerometer data and lowering the duty cycle of voice classifier. The accelerometer data is less useful provided that
we have already obtained the activity inferences (i.e., stationary, walking, running, cycling, in vehicle). The new microphone duty cycle is 1 minute (if no conversation detected) up to 3 (if conversations detected) minutes on and 9 minutes off. Based on these adjustments, we have significantly improved the energy efficiency of the sensing system to support long-term studies. The new system is now being used in on-going 4-year study. In the new study, the participant are satisfied with the battery consumption. We believe we can better understand the trade-off between the accuracy and resources usage as the new project progresses.

While our results are statistically significant and encouraging they are limited to students at UT Austin. The length of the study is only 14 days. We acknowledge this is the first step in this area of research and that more is needed to push forward our understanding of the importance of within-person variability. We welcome researchers to use our within-person features in their studies, and encourage others to conduct similar studies at different sites with different populations to examine whether the findings replicate and are generalizable.

2.7 Conclusion and Future Work

Personality traits describe people’s characteristic patterns of thinking, feeling, and behaving. As such, personality traits describe patterns of variability ‘between-persons’ - that is, patterns of behavior that distinguish people from one another. Personality states, on the other hand, refer to patterns of variability ‘within-persons’. Within-person variability describes fluctuations in how a person’s thinking, feeling, and behaving changes over time. This research reports a mobile sensing approach to assess within-person behavior variability, and thus explores how within-person variability
2.7 Conclusion and Future Work

patterns can be used to predict personality traits using features derived from mobile sensing. Past research has shown within-person behavior variability to be linked to various psychological characteristics. However, much of the past research relies on a person’s capacity to accurately recall their daily experiences retrospectively. Although some researchers have utilized information from smartphones in personality prediction, most work focuses on between-person variability and usually only includes Android participants. Other approaches that use texts on social networks for personality prediction require access to extensive information about people’s online social networks. We designed and implemented a cross-platform mobile sensing study to capture the within-person variability among college students in their social interactions, mobility and movement, daily activities and situational/environmental information. We demonstrate how within-person variability patterns measured by smartphone sensing are related to and thus can be used to predict self-reported personality traits. Our results show that our proposed personality prediction model based on within-person variability features provides good estimation of personality traits, particularly for extraversion and agreeableness. This is, to the best of our knowledge, the first scaled study investigating how within-person variability is predictive of personality traits. It complements and extends existing methods, providing researchers with an additional measure that assesses large groups of participants with minimal burden.

This work represents an important first step toward passive personality assessment. There is a need for the community interested in personality prediction to take the next step and conduct a large scale, longitudinal study with a diverse cohort (e.g., including students, working adults, the elderly). As a contribution, our system can be
2.7 Conclusion and Future Work

easily deployed to collect the necessary mobile sensing data for behavioral tracking in new personality-related projects. It has some potential applications. First, a hybrid approach that combines mobile sensing data with the content available on social networking such as Facebook and Twitter would likely improve predictive performance. Second, the sensing system offers a practical alternative for passive personality assessment, allowing assessment of psychological characteristics in large-scale applications when questionnaires are impractical. The goal of such an approach is to achieve personality assessment without any human intervention. Such an assessment technique could be widely used in recommendation systems, recruiting procedure, for target filtering and for many human-centered applications. Finally, using mobile sensing for measuring within-person variability in behavioral patterns is not limited to personality. This method can be adapted for use in other research areas, such as those focused on prediction of psychological well-being, mental health, or even workplace performance.
Chapter 3

Social Sensing: Assessing Social Functioning of Patients Living with Schizophrenia using Mobile Phone Sensing

3.1 Introduction

Social functioning describes how well an individual is able to interact with their environment and fulfill key skills associated with the social roles they hold within the environment, such as in the domains of social activities, work, and family relationships [36, 246, 326, 319, 268, 264, 143]. As such, social functioning is a key factor of interest in many research areas aimed at understanding and assessing mental health. For example, social functioning has been a key factor in understanding the
3.1 Introduction

experiences of individuals suffering from schizophrenia-spectrum disorders [131, 118], depression [36, 355], and personality disorders [255]. Beyond the domain of mental disorders, deterioration in social functioning is also associated with the aging process because elderly people can also experience a reduced capacity and vulnerability to stress [358, 31, 177, 221] that can impact their ability to successfully interact with their environment and perform their social roles on a day-to-day basis. Given the significance of social functioning as a key factor in overall well-being, there is growing interest in developing passive assessments for monitoring social behaviors, measuring the results of therapy, and designing accurate interventions. However, the standard practice for assessing social functioning in clinical and non-clinical settings to date are typically based on self-reports (e.g., survey responses) or face-to-face evaluations (e.g., interviews). While informative, such approaches are often costly, labor-intensive, and suffer from limited ecological validity [328]. What is needed are new approaches to passively and continuously assessing social functioning in the context of daily life so that timely interventions can be designed. Inspired by this need and other emerging studies on how mobile technologies can be used to understand and assess mental health [53, 15, 342, 193, 288, 287, 26, 23, 348, 80], in this chapter we ask: (1) What behavioral patterns are associated with social functioning in daily life? (2) Can we predict social functioning through passive behavioral sensing assessments derived from people’s smartphones?

To answer these questions, we collected a rich and real-world dataset of social functioning measures and multimodal mobile sensing data from people living with schizophrenia, who were outpatients at a large psychiatric hospital in New York City. Impairment in social functioning is a central feature of schizophrenia symptoms that
3.1 Introduction

is present in most patients [132, 90, 71]. The importance of social functioning in the assessment of schizophrenia has been acknowledged in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5 [13]); the standard classification system for mental disorders used by mental health professionals in the United States. People who have been clinically diagnosed with schizophrenia tend to vacillate between periods of remission and episodes of symptom exacerbation [317], which often triggers shifts in behavioral patterns. These symptom features require the collection of a high-quality longitudinal dataset of social functioning in daily life to capture the considerable variability and change in the behavioral patterns.

The mental health sensing study, CrossCheck, was first introduced in [24]. To date, researchers have leveraged the dataset for innovative analyses focused on assessing problematic symptoms that indicate relapse in schizophrenic patients. For example, Wang et al. [341] developed inference models based on sensing data to predict symptoms that were self-reported from patients (e.g., seeing things, hearing voices, feeling depressed) with a mean error of 7.6% of the score range. In [346], passive sensing data was used to predict Brief Psychiatry Rating Scale (BPRS) scores, which is a 7-item measure that describes symptoms directly associated with schizophrenia that are provided by clinicians (not the patients themselves) with ±1.59 MAE. However, these studies were narrowly focused on monitoring the psychotic symptoms themselves (e.g., delusions, hallucinations, disorganized speech, catatonic behavior [47, 13]) to identify whether an individual was at an increasing risk of schizophrenic relapse [24]. But the findings from this previous work are less sensitive to understanding the general social functioning of these patients.

In this chapter, we focus on understanding and assessing the everyday behavioral
patterns of schizophrenic patients in order to predict their general social functioning. Differentiated from the work above that focuses on schizophrenic symptoms or risk of relapse, we specifically investigate the social functioning, a more generalized area of broader interest to the HCI community. We combine our year-long mobile phone sensing data collected in the wild with detailed social functioning measurements clinically administered every 3 months. Social functioning is measured by the *Social Functioning Scale (SFS)* [33] questionnaire, which rates social interactions, interpersonal relationships, and activities of independent living in 7 sub-areas: social engagement/withdrawal, interpersonal behavior, prosocial activities, recreation, independence-competence, independence-performance and employment/occupation. We find the impairments and disabilities in each detailed area of social functioning are associated with different behavioral signals captured by smartphones. Our approach is an instrumental step toward designing timely interventions that can be deployed in the context of people’s natural lives, and can lead to behavioral insights that generalize to other populations with mental health issues or other social functioning difficulties.

The structure of the chapter is as follows. We start by discussing the related work on social functioning and then detail our study, dataset, models, methods, results and insights. We also discuss the privacy and ethical implications surrounding the collection of sensitive data from populations at risk – we make some concluding remarks on this important topic in the discussion section. Our results show mobile phones can accurately predict social functioning impairment among the people living with schizophrenia in our study. Our results on sensing social functioning open the way to new forms of assessment and intervention across a number of areas important
3.1 Introduction

to the HCI community, such as, mental health, aging in place, cognitive decline and dementia.

3.1.1 Privacy, Ethics and Disclosure

While it is clear that clinicians and patients could benefit from passive sensing, there are risks associated with collecting sensitive data. Considering the broader concern about ethics and privacy in this field, the misuse of such technologies could cause serious privacy issues. In what follows, we discuss how such sensitive sensed data and inferences should be protected.

First, researchers must make sure that participants knowingly consent to give up some of their privacy during the study. We explicitly stated the risks of participation in our consent forms, and discussed this in detail while enrolling participants. During enrollment, participants also had the opportunity to ask questions and were given supportive materials including visuals outlining all sensors and information collected by the app. Prior to obtaining consent, we also tested participants’ knowledge of the type of data collected, whether the study was confidential, not anonymous, and ways in which we protect their data. Participants were administered a competency screener to verify that they understood the details and were able to provide informed consent. They were not allowed to participate if they could not pass this test.

We have taken great effort to protect the privacy of our participants. We took extra precautions with this vulnerable population when developing the study, software and refined the data we collected down to only the essential identifiable sensors and data captured for this study. Almost all the data captured by the app was unidentifiable. For example, we developed unique processing (conversation detecting
algorithms) on the device itself to capture the presence of socialization, omitting the need to capture any actual spoken word. The voice detection algorithm only detects whether the person is around speech, and does not collect the content of conversations, only the frequency and duration of conversations. We also captured light sensing, accelerometer, and the number of calls and texts a participant sends during the course of a day; each unidentifiable, and if disclosed, would have little consequence to the participant. We did not collect other types of sensitive information (e.g., web-browsing, keyboard strokes, social media content, friends/family’s phone numbers, pictures, videos). In addition, all the sensed data is decoupled from demographic data and is associated with a random study ID.

3.2 Related Work

HCI researchers have examined ways of using passively collected sensing data to infer individual and community behaviors and mental states in different contexts [293, 138, 151, 239]. For example, Alharbi et al. [7] designed WillSense, a wearable device to collect fine-grain eating behaviors in the wild among people with obese, overweight, and normal BMI. Li et al. [201] utilized IDSense tags for unobtrusive human object interaction detection that enables inferring daily activities at home. Sun et al. [320] used MoveMeant, a location awareness app to monitor local community ties and support local community building. HCI researchers have also leveraged naturalistic content from online communities to address broad topics including self-disclosure [216, 368], behavioral prediction [58, 59], community norms [57] and mental illness [100, 108].

Mental health is one of the common domains underlying these efforts. In recent
years, there has been increasingly influential work on mobile sensing for mental health [53, 15, 342, 193, 288, 287, 26, 23]. In [273], the authors used an early mobile sensing platform device [66] to show how conversation and physical activity can be used to infer mental and social well-being among elderly adults living in a continuing care retirement community. The StudentLife study [342] demonstrated that depression, stress, loneliness, and flourishing are associated with passive sensing behaviors (e.g., conversation, sleep, activity, co-location). Seab et al. [288] investigated that location features extracted from GPS data including circadian movement, normalized entropy, and location variance, and phone usage features including usage duration and usage frequency are associated with depressive symptoms. The findings from their initial study were replicated [287] using the StudentLife dataset [342]. Canzian et al [53] developed an extended set of mobility features over [288, 287] to show how location data is related to depression [314, 187, 186]. Wang et al. [347] proposed a set of symptom features derived from phone and wearable sensors that proxy the DSM-5 defined depression symptoms specifically designed for college students. Sarker et al. [293] propose a pattern mining method to detect significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. Researchers also develop remote monitoring tools for clinicians or self-care interventions for mental health. Schroeder et al. [296] contribute a conversational mobile web app to help people with complex disorders maintain positive relationships, and control their emotions. The intersection between ubiquitous computing and sensing, social media and emerging technologies offers promising avenues for novel human-centered designs in enhancing mental wellbeing [337].

Social functioning impairment in schizophrenia [47] is studied widely by sev-
eral authors. The studies range from understanding human social cognition which is highly affected by schizophrenia to studying how it is impaired by the illness. The work in [33] discusses new methods for quantifying social functioning impairment in schizophrenia using the SFS survey to assess social functioning. Other researchers [21] show that social cognitive impairments can be mitigated with the help of family members of patients with schizophrenia by being conscious of their expressed emotion. HCI researchers have studied social functioning impairment across a wider range of conditions; for example, aging, cognitive decline and dementia [358, 31, 177, 221, 358, 208, 55], as well as hearing impairment [160] and post-traumatic stress disorder [107]. HCI researchers have designed or evaluated innovative systems that offer social support and improve social interaction. Adams et al. [4] examine the health applications of staccato social support in mobile environments designed for brief, rapid social sharing and interaction. Wiley et al. [358] design Message Center, a home-based communication solution for enhancing elder communication. Laput et al. [208] propose the StoryCubes system that creates mutual understanding and appreciation between independent living residents through the experience of telling and listening to stories.

3.3 Dataset

The CrossCheck study [341] is a randomized control trial (RCT) [56] conducted in collaboration with a large psychiatric hospital, Zucker Hillside Hospital, in Long Island, New York\textsuperscript{1}. The whole study lasted for 4 years and recruited 150 participants.

\textsuperscript{1}This study was approved by the Committee for Protection of Human Subjects at Dartmouth College and Institutional Review Board at North Shore-Long Island Jewish Health System.
3.3 Dataset

for 12 months using rolling enrollment. The participants are randomized into either the smartphone arm (n=75) or treatment-as-usual arm (n=75). Potential study candidates are identified according to the hospital’s electronic medical records [24, 341]. Participants receive Samsung Galaxy S5 Android phone equipped with the CrossCheck app and receive a tutorial on how to use the phone. They are asked to keep the phone turned on, to carry it with them as they go about their day, charge it close to where they sleep at night and also answer EMA every 2-3 days. The research staff check the daily report of sensed data and would call noncompliant patients to assist and get them back on track\(^2\). Due to these efforts, the sensing data shows good compliance except for the cases when the participants had relapses and were hospitalized with the phones being off. We exclude these days during the analysis (thus all the aggregated values discussed in the chapter are normalized by the number of valid data during a time frame). In this chapter, 55 participants in the smartphone arm who were complying with the study and social functioning assessment are involved in the analysis. Table 4.1 shows demographics of 55 participants. The demographics in our sample is also reflective in the ratio reported in other research [42].

3.3.1 Mobile Sensing Data

The CrossCheck app collected a wide range of behavioral passive sensing data from the phone. The details of system design have been discussed in the prior work [341, 346]. More specifically, it captured physical activities [125], locations [129], ambient

\(^2\)To ensure the acquired data has a broad coverage of behaviors, participants’ personal phone numbers are migrated to the new phone and they are provided with an unlimited data plan for data uploading. In addition, to encourage the adherence, participants are compensated $20 every 3 months during the assessment based on participation and the length they were in the study. We also give them the phone after the completion of the study.
3.3 Dataset

<table>
<thead>
<tr>
<th>race</th>
<th>count</th>
<th>percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian</td>
<td>1</td>
<td>1.82%</td>
</tr>
<tr>
<td>Native Hawaiian or Other Pacific Islander</td>
<td>3</td>
<td>5.45%</td>
</tr>
<tr>
<td>Black or African American</td>
<td>17</td>
<td>30.91%</td>
</tr>
<tr>
<td>White</td>
<td>19</td>
<td>34.55%</td>
</tr>
<tr>
<td>More than one race</td>
<td>13</td>
<td>23.64%</td>
</tr>
<tr>
<td>Unknown or Not Reported</td>
<td>2</td>
<td>3.64%</td>
</tr>
</tbody>
</table>

Table 3.1: Demographics of participants

sound levels [128], voice/noise labels [273], number of calls and text messages [126], application usage, screen lock/unlock, and ambient light intensity [127]. We compute features from the passive sensing data on a daily basis, which describe participant’s behaviors (e.g., duration of different physical activities in a day, conversation duration and frequency, different types of places visited, app usage).

3.3.2 Social Functioning Assessment

A clinical assessor administers each participant’s symptom severity, depression, and social functioning in person every 3 months during the year-long study. As discussed earlier we use the SFS survey [33] to measure the social functioning of outpatients in our study. The questions of SFS are informed by the Disability Assessment Schedule [159] and previously successful psychosocial intervention programs. SFS is shown to be valid, reliable, and sensitive to assessing a range of social functioning impairment [33]. As discussed SFS consists of 7 sub-scales: 1) social engagement/withdrawal, which measures time spent alone, initiation of conversation and social avoidance; 2) interpersonal behavior, which captures number of friends/having a romantic partner and quality of communication; 3) prosocial activities, which assesses the engagement...
3.3 Dataset

in a range of social activities (e.g., sport); 4) recreation, which gauges the engagement in a range of hobbies, interests, and pastimes; 5) independence-competence, which inquires about the ability to perform skills necessary for independent living; 6) independence-performance, which rates the performance of skills necessary for independent living; and finally 7) employment/occupation, which relates to the engagement in productive employment or a structured program of daily activity. A higher sub-scale score indicates better social functioning.

Table 3.2: Statistics of Social Functioning Scale

<table>
<thead>
<tr>
<th>SFS sub-scale (scale range)</th>
<th>response range</th>
<th>overall mean(std)</th>
<th>within-person range*</th>
</tr>
</thead>
<tbody>
<tr>
<td>social engagement (0-15)</td>
<td>2-15</td>
<td>11(2.5)</td>
<td>3(2-5)</td>
</tr>
<tr>
<td>interpersonal behavior (0-30)</td>
<td>4-29</td>
<td>20.8(4.9)</td>
<td>5(3-8)</td>
</tr>
<tr>
<td>prosocial activities (0-66)</td>
<td>0-61</td>
<td>20(12.2)</td>
<td>12(8-17)</td>
</tr>
<tr>
<td>recreation (0-48)</td>
<td>0-39</td>
<td>20.6(8.5)</td>
<td>10(5-16)</td>
</tr>
<tr>
<td>independence-competence (0-39)</td>
<td>19-39</td>
<td>36.7(3.4)</td>
<td>3(1-5)</td>
</tr>
<tr>
<td>independence-performance (0-39)</td>
<td>7-39</td>
<td>28.3(6.7)</td>
<td>7(5-10)</td>
</tr>
<tr>
<td>employment (0-10)</td>
<td>0-10</td>
<td>5.3(3.4)</td>
<td>2(1-4)</td>
</tr>
</tbody>
</table>

* Values in median(LQ-UQ) where LQ and UQ means the lower quartile (25%) and upper quartile (75%) respectively.

Table 3.2 shows the range of the sub-scales, the range of the scored responses from our study participants, the overall mean and standard deviation and the median, lower and upper quartile of the ranges of within-person changes among the multiple assessments during the year. Many participants experienced varying levels of social functioning during the study, which consents to other longitudinal studies in social functioning [334, 166, 89, 5]. We also observe that a participant may score higher in certain sub-scales and lower in others. For example, one participant has higher scores in social engagement and interpersonal behavior but lower scores in prosocial activities.
3.4 Model and Method

3.4.1 Behavioral Features

We incorporate features that describe the six aspects of daily behaviors from the smartphone passive sensing data. Specifically, we compute the features of 8 categories listed in Table 5.3, including physical activities, mobility, sleep patterns, ambient environmental context, face-to-face conversations, smartphone-based communications, smartphone usage and semantic location. Features are computed on a daily basis and broken down into four epochs of the day: morning (6am-12pm), afternoon (12pm-6pm), evening (6pm-12am) and night (12am-6am), that allow us to model people’s behaviors during different parts of the day. Note that we compute new features for social functioning that has not been well-studied before. For example, smartphone-based communications features now include the usage of apps for voice or video calls (e.g., Skype and Hangouts); the time on different types of apps, e.g., social networking (Facebook, etc.), game, browser, entertainment (music/audio/video, etc.), and engagement (finance/tools/education/business, etc.) is also well investigated. Importantly, we consider semantic location features. SFS contains questions about activities at different types of places. For example, the “Social Engagement/Withdrawal” section asks “do you leave the house”; the “Independence-Performance” section asks about “buying everyday items from stores”; and the “prosocial” section asks about “visiting art gallery/museum”. Motivated by the questions, we aim to assign semantics to places where participants visit. Specifically, we consider the following places: home, food, travel, art & entertainment, nightlife, education, park & outdoors, library, shop, gym, medical and residence. We compute the time spent at these places.
3.4 Model and Method

during a certain time frame.

Table 3.3: Features computed from mobile sensing data. We have 115 behavioral features in total. The length of each feature is 160, corresponding to the number of SFS responses adopted in our dataset. The unit could be in seconds (e.g., duration of various activities, conversations, etc.), an integer of counts (number of places visited, number of lock/unlocks, etc.), or ambient light level/ambient sound amplitude directly from sensors.

<table>
<thead>
<tr>
<th>category</th>
<th>details</th>
</tr>
</thead>
<tbody>
<tr>
<td>physical activities</td>
<td>duration on foot / in vehicle / on bicycle / sedentary</td>
</tr>
<tr>
<td>mobility</td>
<td># of locations visited, distance travelled</td>
</tr>
<tr>
<td>sleep patterns [63]</td>
<td>sleep duration, sleep start time, and sleep end time</td>
</tr>
<tr>
<td>ambient env. context</td>
<td>amplitude of ambient sound, level of ambient light</td>
</tr>
<tr>
<td>f2f conversations</td>
<td># of conversations, duration of conversations</td>
</tr>
<tr>
<td>smartphone-based communications</td>
<td># in-coming phone calls, # of out-going phone calls, # of in-coming SMS, # of out-going SMS, duration of usage of apps for voice or video calls</td>
</tr>
<tr>
<td>smartphone usage</td>
<td># of lock/unlocks, unlocked duration, duration of using different genres of apps: social networking, game, browser, entertainment and engagement.</td>
</tr>
<tr>
<td>semantic location</td>
<td>duration @ home, food, travel, art&amp;entertainment, nightlife, education, parks&amp;outdoors, library, shop, gym, medical and residence.</td>
</tr>
</tbody>
</table>

To label locations with semantics, we first cluster the raw GPS coordinates using density-based spatial clustering of applications with noise (DBSCAN) [101]. The centroid of each cluster is considered a significant location, where a participant dwells for a significant amount of time. We first label a significant location as the home where a participant spends most of the time and typically frequents between 2 am to
3.4 Model and Method

6 am. We use the Foursquare API [112] to label the other significant locations. It takes a GPS coordinate and a radius as input and returns a list of location entities. Each location entity is associated with a name, coordinates, and categories (e.g., food, art&entertainment). We test the sensitivity based on different radiiuses: 50m, 30m and 20m. Overall, 50m outperforms the rest in predicting the social functioning scores. Therefore, we use the semantic features from a radius of 50m in the analysis. A location may be associated with multiple different categories. We compute the dwell duration at a location for all associated location categories. For instance, if the API returns “food” and “art&entertainment” categories for a given significant location, we include the dwell duration at this significant location to both “food” and “art&entertainment” places. This is an approximation given that there is an error in location coordinates. We do not simply select a single returned location entity closest to the significant location coordinate because as mentioned GPS data is noisy [98] and the same building could be associated with different semantics.

3.4.2 Association Analysis

In order to understand how the passive sensing data is associated with SFS, we use principal components analysis (PCA) [295] and generalized estimating equations (GEE) [202, 145] to explore the associations between the principal components of behavioral features computed from passive sensing data and SFS sub-scale scores. Specifically, we look at the sensing data within 90 days prior to the SFS response. In what follows, we discuss our association analysis method in detail.

**Principal Components Analysis.** We compute the mean for each daily feature (e.g., mean of daily conversation duration in the 90-day time frame) per participant.
3.4 Model and Method

After computing the features, we organize the features in a matrix $M$, where a column represents one of the 115 features, and a row represents features that are associated with one of the 160 SFS responses. We apply PCA on $M$ to find a small number of principal components (PCs) that explain most of the variance in the feature matrix $M$. Using PC scores in our analysis presents several advantages over using behavioral features. First, selecting a smaller number of PCs significantly reduces the feature dimensionality. Second, each PC can be interpreted as a behavioral pattern. For example, if a PC has large positive weight in the component for phone unlock duration and phone call duration features, and large negative weight for still duration, we would interpret this PC represents a high phone usage and high sedentary behavioral pattern. The absolute value of the PC score represents how significant this pattern is to a row in $M$ (i.e., a participant’s behavioral patterns within the 90 days prior to the SFS response). Figure 3.1 shows the cumulative variability in the data explained by the top-n principal components. We select the top 23 PCs in $M$ that explain 90% of the variance in our analysis.

![Figure 3.1: The cumulative variability in the data explained by the top-n principal components. We include the top 23 principal components that explain 90% of the variance in our analysis.](image)
of the variance for our analysis.

**Bivariate GEE.** We apply bivariate GEE [202, 145]\(^3\) to investigate how the selected 23 PCs are related to SFS scores by regressing the PC scores to SFS sub-scale scores. The SFS data is longitudinal: we receive multiple responses from one participant at different time in the year long RCT study. The responses from the same subject are dependent. We apply GEE to the combinations of 23 PCs and 7 SFS scores. To determine the false discovery rate (FDR), we perform the two-stage Benjamini-Hochberg procedure (TSBH) [29]\(^4\).

### 3.4.3 Predicting Social Functioning Scale

We experiment with different methods to find the best predictive model for social functioning scale. Specifically, we experiment with different time windows, different dimensionality reduction methods and different machine learning models. In what follows, we discuss our method in detail.

**Prediction Time Window.** Social Functioning Scale asks a respondent’s experience over the last 3 months. Therefore, it is intuitive to use the sensing data within 90 days prior to the response to predict the SFS scores. However, we suspect participants’

---

\(^3\)The GEE method is designed to analyze longitudinal data. It is an extension of Generalized Linear models (GLM) [227] to model the regression and within-subject correlation separately. The GEE approach fits marginal models, therefore is commonly used to estimate population-averaged effects. Suppose \(Y_{it}\) represents the response from participant \(i\) at the \(t\)th measurement, the bivariate GEE model is usually described as
\[
g(\mu_{it}) = \beta_0 + \beta x_{it},
\]
where \(\mu_{it} = E(Y_{it})\) denotes the expectation of the response for subject \(i\) at time \(t\), \(x_{it}\) is the value of feature for participant \(i\) at time \(t\) from sensing data, \(g\) indicates a link function, \(\beta_0\) is the intercept, and \(\beta\) denotes the association between the sensing feature and the response. The p-value associated with \(\beta\) shows the probability of the coefficient being zero.

\(^4\)The TSBH first examines the distribution of p-values to estimate the fraction of the null hypotheses that are actually true. It then uses this information to decide when a p-value is low enough to be called a discovery. This presents a rigorous approach to the analysis.
3.4 Model and Method

responses may subject to recall bias such that the responses may be more accurate to
describe the participants’ recent behaviors. Therefore, we also examine the predicting
performance using shorter time windows. We assess the prediction performance with
90-day, 60-day, 30-day, and 15-day time windows.

**Feature Transformation and Dimensionality Reduction.** We experiment three
different approaches for feature transformation and dimensionality reduction: PCA
[360], truncated SVD [137], and Kernel PCA [295] with radial basis function kernel.
We assess the prediction performance of using different number of principal compo-
nents explaining 70%, 80%, 90%, and 95% of the variance for different time windows.

**Evaluation of Machine Learning Models.** We evaluate the prediction perfor-
mane of different time windows, dimensionality reduction methods, and prediction
models. We evaluate the prediction performance of linear regression (Lasso
[324]/Ridge [154]/ElasticNet [377]), linear support vector regression (LinearSVR)
[93], support vector regression with the radial basis function kernel (SVR_rbf) [309],
extremely randomized trees (ET, Extra Trees) [121], random forest (RF) [41], and
XGBoost [61].

We first apply 5-fold cross-validation to evaluate the prediction performance of a
time window, dimensionality reduction method, and predictive model combination.
We first iteratively split the data into 80% training and 20% hold-out testing. We then
use the training fold to tune model hyper-parameters (e.g., regularization strength,
learning rate, depth of regression trees, number of estimators) which maximizes pre-
diction performance using 5-fold cross-validation. We train the predictive model with
the best hyper-parameter using all the training data and test on the test set. We re-
port test mean absolute error (MAE), root-mean-square error (RMSE) and Pearson
correlation between predicted values and the SFS ground truth.

During the training process, in some folds, the data from the same subject could be split to both training and testing groups while the data of other subjects might only exist in either training or testing group. Therefore, we consider these mixed settings as where some of the patients may have been assessed multiple times while others may not. For example, there are some patients who were hospitalized and had been assessed, however, clinicians must still keep track of their status even after they are discharged. Nevertheless, there are also other patients who do not have any SFS records in the system, and clinicians need to get an initial sense of their status. The 5-fold cross-validation training process reports averaged performance metrics of models considering these circumstances.

To better understand the robustness of the system, we further perform the leave-one-subject-out (LOSO) validation. In the LOSO, models are trained on the data from other study participants with one subject’s data left out. This emulates a new patient without any historical SFS records. In addition, we perform a customized leave-one-subject-out (CLOSO), which uses the first data point (in chronological order) of one specific patient combined with all the data points of other patients to predict the responses of that patient in latter assessments. This corresponds to a situation (which happens mostly in real practice) that patients have established their first medical record in the hospital and will rely on the app to track the social functioning afterwards.
3.5 Results

In what follows, we present results for association analysis and prediction of social functioning.

3.5.1 Association Analysis

We investigate the correlation between the PC scores associated with 23 PCs and the SFS sub-scale scores using GEE and FDR as discussed earlier. We label each principal component by investigating its loadings (i.e., the weights associated with features). We rely on the highly ranked behavioral features (i.e., features with larger absolute weight in the component) to understand the typical behaviors that make up each principal component. Our results show that 10 of the 23 behavioral patterns (i.e., principal components) are associated with various social functioning sub-scales. We list the 10 behavioral patterns and their corresponding features in Table 3.4, where the features in each pattern are ordered by the absolute value of the weight.

Table 3.4: Patterns associated with 7 sub-areas of social functioning

<table>
<thead>
<tr>
<th>SFS sub-scale</th>
<th>associated behavioral pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>social engagement</td>
<td><em>pattern 7</em> (-), pattern 17 (+)</td>
</tr>
<tr>
<td>interpersonal behavior</td>
<td>pattern 1 (+), pattern 14 (-)</td>
</tr>
<tr>
<td>prosocial activities</td>
<td>pattern 1 (+)</td>
</tr>
<tr>
<td>recreation</td>
<td>pattern 11 (-), pattern 21 (-), pattern 3 (+)</td>
</tr>
<tr>
<td>independence-competence</td>
<td>pattern 9 (+), pattern 4 (-)</td>
</tr>
<tr>
<td>independence-performance</td>
<td><em>pattern 1</em> (+)</td>
</tr>
<tr>
<td>employment</td>
<td><em>pattern 3</em> (+), pattern 13 (+)</td>
</tr>
</tbody>
</table>

(-) negative association; (+) positive association, all associations with $p < 0.05$. $FDR < 0.1$ in **bold** and $FDR < 0.05$ in **bold italic**
3.5 Results

**Behavioral Patterns.** The behavioral patterns are listed in Table 3.4 in detail. It shows the behavioral characteristics of each pattern indexed by the n-th component of PCA (e.g., Pattern 1 where 1 is the first component of the PCA). The listed patterns are associated with social functioning sub-scales. Specifically, Pattern 1 describes people who often talk on the phone; Pattern 3 describes people who travel longer distance during evening and night (often using vehicle); Pattern 4 describes people who communicate with others using SMS rather than face to face social interactions captured by the audio sensor periods; Pattern 7 suggests people who are more active on foot, spend more time in brighter places in the mornings and afternoons periods, and have less exposure to face to face conversation interaction – more socially isolated; Pattern 9 describes people who do more exercise on bikes; Pattern 11 is associated with people who like to Spend time in educational settings and sleep late and be gamingless; Pattern 14 describes individuals who show up frequently in libraries, show up less in residential areas (e.g., places which may be the homes of their friends) but not in the residential area where they live; Pattern 17 describes people who tend to rarely visit the gym, go to bed early, rarely make long phone calls during the night, spend less time at home, use more communication apps (e.g., Hangout, Skype, Messenger), and use less social networking apps (e.g., Facebook; pattern 21 suggests people who are mostly engaged in playing games on the phones.

**Associations between behavioral patterns and social functioning sub-scales.** Table 3.4 shows the behavioral patterns that are associated with the 7 social functioning sub-scales. A higher sub-scale score indicates better abilities and performance in a social functioning domain. A positive association between a pattern and a sub-scale suggests people whose behavior matches such pattern are more likely to have a higher
3.5 Results

sub-scale score, thus higher ability. Conversely, a negative association suggests people whose behavior matches such pattern are more likely to have a lower score (i.e., more likely to have impairment). In what follows, we describe our findings.

Social engagement sub-scale measures time spent alone, initiation of conversation, and social avoidance tendencies. It asks questions about the time one gets up, the hours one spends alone, the frequency one starts a conversation at home, the frequency one leaves home, and how one reacts to the presence of strangers. We find the social engagement sub-scale is negatively associated with pattern 7 (being physically active outside during the day without others around) and positively associated with pattern 17 (spending less time in the gym and going to bed early).

Interpersonal behavior sub-scale asks about the number of friends, the quality of the relationship, and communications the person has with friends (e.g., “do people discuss their problems with you?”, “how often are you able to carry out a sensible or natural conversation?”). We find higher interpersonal behavior scores are positively associated with pattern 1 (using the phone frequently for phone calls) and negatively associated with pattern 14 (spending time in the library and visiting few places in the evenings).

Prosocial activities sub-scale contains 22 questions measuring detailed social activities in everyday life. The questions ask how often people go to the movies/theatre/concert, watch/play indoor/outdoor sports, visit art gallery/museum/exhibition/fair, visit relatives/friends, go to pub/bar, eat out in restaurant, etc. We find that pattern 1 (using the phone frequently for phone calls) is positively associated with higher scores in the prosocial activity sub-scale.

Recreation sub-scale asks people’s engagement in a range of hobbies and inter-
3.5 Results

ests, such as singing, playing instruments, reading, gardening, watching television, cooking, and hiking. Pattern 11 (spending time in educational settings and sleeping later and less gaming) and pattern 21 (merely using gaming apps and going to sleep late) are negatively associated with recreation score whereas pattern 3 (traveling and socializing with others at night) is positively associated with recreation. The finding suggests participants who are engaged in educational settings or mobile games tend to have fewer hobbies and interests, while those who socialize with other people during the night are more likely to enjoy various recreations.

*Independence-competence* sub-scale measures people’s ability to perform the skills necessary for living independently. Such abilities include taking public transportation, budgeting, handling money correctly, cleaning, weekly purchasing, taking care of personal appearance. These abilities are usually internal and hard to be captured by passive sensing. However, we find two patterns that indicate the strength and weakness in this scale. Stronger pattern 9 (bike riding and spending time in different places) indicates a higher independence competence score whereas weaker pattern 4 (texting and chatting on the phone without others around) indicates a lower independence competence score.

*Independence-performance* sub-scale examines the performance of skills necessary for independent living. Unlike the Independence-Competence sub-scale, the independence-performance sub-scale asks how often people perform activities necessary in independent living rather than their self-evaluated abilities. We find that pattern 1 (talker over phone) is positively associated with the performance in activities. Although calling and texting are often associated with interpersonal motives [171], no prior work shows the correlations between using calling on phones and the
3.5 Results

performance of skills necessary for independent living. There is a need to better understand the reason for such connections.

*Employment or occupation* sub-scale measures engagement in employment or structured daily activity program. The score ranges from 0 to 10 where 10 indicates either full time gainful earnings, full time employment, or are homemaker managing most household affairs for self and 0 indicates unable to make attempts to find a job [33]. Pattern 3 (traveling and socializing with others at night) and pattern 13 (Sleeping and waking earlier) are positively associated with engagement in employment or occupation. Pattern 3 describes people who are likely to enjoy the nightlife whereas pattern 13 portrays who are early to bed, early to rise and having fewer activities on foot during the afternoon and evening. They might depict two styles of employers: some are night owls who like to have fun after work until late, and some prefer early hours. It is not surprising since traveling in a vehicle (regular commute) in the evening and night is very likely to be a sign of employment. On the other hand, existing literature suggests the link between work schedules and chronotype, in which the authors found the unemployed are less likely to be morning types [259].
## 3.5 Results

Table 3.5: Prediction Summary using leave-one-subject-out and customized leave-one-subject-out sub-scale model 5-fold cross validation

<table>
<thead>
<tr>
<th>sub-scale</th>
<th>model</th>
<th>window/feature</th>
<th>5-fold cross validation MAE/RMSE</th>
<th>Pearson r-value</th>
<th>Pearson r-value vs. baseline MAE/RMSE</th>
<th>Pearson per-user error mean(std)</th>
<th>Pearson per-user error mean(std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>social</td>
<td>15 / RF / kernel pca 80% features</td>
<td>0.44</td>
<td>1.75 / 2.28</td>
<td>0.44</td>
<td>1.75 / 2.28</td>
<td>-0.12 / -0.28</td>
<td>0.22</td>
</tr>
<tr>
<td>interpersonal behaviors</td>
<td>60 / ET / orig. features</td>
<td>0.57</td>
<td>3.39 / 4.26</td>
<td>0.57</td>
<td>3.39 / 4.26</td>
<td>-0.56 / -0.88</td>
<td>0.38</td>
</tr>
<tr>
<td>prosocial activities</td>
<td>90 / XGB / orig. features</td>
<td>0.53</td>
<td>7.79 / 10.35</td>
<td>0.53</td>
<td>7.79 / 10.35</td>
<td>-2.11 / -1.91</td>
<td>0.23</td>
</tr>
<tr>
<td>recreation</td>
<td>15 / XGB / orig. features</td>
<td>0.39</td>
<td>6.62 / 8.15</td>
<td>0.39</td>
<td>6.62 / 8.15</td>
<td>-0.72 / -0.64</td>
<td>0.16</td>
</tr>
<tr>
<td>independence competence</td>
<td>90 / ET / kernel pca 95% features</td>
<td>0.51</td>
<td>2.06 / 2.91</td>
<td>0.51</td>
<td>2.06 / 2.91</td>
<td>-0.31 / -0.49</td>
<td>0.23</td>
</tr>
<tr>
<td>independence performance</td>
<td>90 / ET / orig. features</td>
<td>0.47</td>
<td>4.49 / 5.78</td>
<td>0.47</td>
<td>4.49 / 5.78</td>
<td>-0.80 / -0.81</td>
<td>0.23</td>
</tr>
<tr>
<td>employment</td>
<td>90 / ET / orig. features</td>
<td>0.62</td>
<td>2.17 / 2.67</td>
<td>0.62</td>
<td>2.17 / 2.67</td>
<td>-0.76 / -0.74</td>
<td>0.38</td>
</tr>
</tbody>
</table>

81
3.5 Results

3.5.2 Predicting Social Functioning Scale

We present the best combinations of time window, feature transformation, and predictive model to predict the SFS sub-scale scores in Table 3.5. The performance is reported using mean absolute error (MAE), root-mean-square error (RMSE), Pearson correlation, the association between the predicted and the true values grouped by 5 folds in cross validation. We also compare the MAE and RMSE of our model with the baseline model, in which every test case is predicted using the mean of the scores from the training data. Overall, the tree-ensemble models (RF, ET & XGBoost) outperform linear regression and SVR models. The mean absolute error is around 10% of the possible range of each sub-scale. The predicted SFS scores are positively associated with their responses. All the predictive models are performing better than the baseline models.

To further understand the performance under various scenarios, we present the accuracy from LOSO and CLOSO. Table 3.5 shows the performance of the best models. In general, we find that the pure LOSO process performs poorly. This indicates that our dataset is still not sufficient enough in order to train a model purely on the data from other people. This is somehow expected; social functioning sub-scales consist of many items that measure a specific type of functioning from multiple behaviors. Those behaviors could be highly personalized, even among people with similar scores. Behaviors, unfamiliar to the model from an unseen user, prevent the system to be trained well. Therefore, we still need to leverage knowledge from the population and a small amount of subjective behavioral patterns; in this case, we adopt the first chronological response of the participant. This additional data point serves as an initial anchor; the models then predict the following variances based
3.5 Results

on the knowledge from the population. The performance of CLOSO is significantly better than LOSO’s and close to 5-fold cross validation’s. We calculate the mean absolute error for each participant during the LOSO and CLOSO, and thus compute the distribution of the per-user errors. The per-user absolute errors from CLOSO are less than 15% of the total ranges for most of the sub-scales.

Previous studies in sensing schizophrenia show that the variations or trajectory of sensed behaviors are informative in predicting the dynamic symptoms. However, including these features did not show any significant improvement in our prediction. We assume these features are unimportant to prediction social function where it asks about the recent/past aggregated activities. Therefore, we choose a simple approach; that is, we adopt only the average of sensed behaviors in d days (where d = 90, 60, 30, 15) before the date of the assessments as predictors in our model. Considering the different characters in sub-scales, SFS measures behavior in various periods for different sub-scales. In particular, participants are asked “how often you have done the following over the past 3 months” in prosocial activities, independence-competence, and independence-performance, while for other sub-scales the period is not explicitly indicated. As shown in Table 3.5, we find different optimum window for each sub-scale. The model selects an optimal window of a 90-day time frame for prosocial activities, independence-competence, and independence-performance, which matches the way social functioning is measured. This verifies that the mobile sensing techniques can capture the daily behaviors for automatic assessment. For sub-scale with questions mostly focusing on current status without specifying a range of time (e.g., social engagement/withdraw and interpersonal communication), we find shorter time frames work better.
3.5 Results

The model performs worst in predicting the recreation sub-scale. The reason might be due to the disability of mobile phone sensors in capturing detailed items for recreation, such as swimming, gardening, cooking, etc. In addition, the authors of SFS [33] report that the reliability ($\alpha$ coefficient $= 0.69$) of the recreation sub-scale is significantly lower than the other sub-scales (0.82-0.87). This indicates that the people could not remember the frequency of various recreational activities over the past 3 months. In future studies, extra mobile sensors can be leveraged for better capturing recreational activities.

3.5.3 Patient Case Studies

Here we highlight one anecdotal case study for showing the effect of this model. The patient in Figure 3.2 is a 50 year-old white female diagnosed with schizophrenia. She has a high school education, is unemployed and lives independently. She was working as a secretary in her last job. She was enrolled in Mar. 2015 and did not experience any relapses during the study. At the beginning of the study, she reported that usually in the morning, she had breakfast, drank coffee, watched news channel, did shower and dress, checked schedule and phoned with friends; in the afternoon, she did job searching, going for work appointment, workout, spending time with husband; and during the evening, she watched TV, then went to sleep. In the first 6 months, she was confident in her capability of employment and was looking for jobs every day. However, her daily occupation significantly changed according to her assessment in Sept. 2015. During that time, she usually slept until 2-3 pm and watched TV with husband during the afternoon. She was still confident that she was capable of some sort of employment, but she rarely did any job hunting. Her last assessment in Mar.
3.5 Results

2016 showed that she watched TV almost all the time in the morning as well as in the afternoon. She reported dinner, cleaning up and sleep as the major occupation in the evening. She thought she would have some difficulty in employment and never looked for a job during that time. Figure 3.2 shows how the CLOSO model predicts her social functioning. The first assessment is included in the training set together with the knowledge from the population.

![Graphs showing social functioning](image)

Figure 3.2: Case study: the dynamics of a participant’s assessed and predicted social functioning using CLOSO. (a) the model identifies the strengths and weaknesses between the different components in social functioning. (b) the model tracks the dynamics of social functioning.
3.6 Discussion

From the predicted values we know: (a) this participant kept high independence competence yet had weakness in prosocial activities during the study and (b) experienced deterioration in her employment sub-scale around Sept. 2015. Although the CLOSO still generates predictions skewed to the first input, the clinicians should be aware of the changes indicated by the model. This user case shows the importance and effectiveness of the proposed system. Limited by the scale of the dataset, we admit the model could be further improved with data from more patients. However, the system is capable of detecting the strengths and weaknesses between the different components in social functioning, as well as tracking the dynamics.

3.6 Discussion

We use a number of computational approaches to better understand behavioral patterns associated with various aspects of social functioning. We demonstrate how we can design features and train a model using mobile sensing data from phones of outpatients with schizophrenia to predict social functioning. To the best of our knowledge, we are the first to use mobile sensing in wild settings to study social functioning. Importantly, our approach in assessing social function is applicable to other areas of interest to the HCI community including, for example, cognitive decline in the elderly and dementia. Furthermore, we believe our work opens the way for the development of new forms of assessment and intervention across these areas.
3.6 Discussion

3.6.1 Implications for Social Computing and Health Research

The computational methods discussed in the chapter augment current methods to improve understanding of social functioning developed by HCI community; for example, face-to-face evaluation [352] (which can be labor-intensive) and extracting behavior and language from social networks, such as Facebook, Twitter and Reddit [46, 301] (which can only captures online behavior and not continuous assessment in the wild afforded by our approach). In the future, an approach that combined mobile sensing and social media may prove to be even more effective.

Our approach highlights the importance of using multi-modal data in assessing various components of social functioning. In addition to features that have proved effective in prior research [341, 346], we show new features (e.g., communications through mobile phone applications like Skype, Hangout, semantic locations) are important signals for social functioning. Specifically, this work demonstrates how researchers can use the classical PCA technique for interpretable and generalizable behavioral patterns associated with different outcomes. Our approach also shows the steps of combining machine learning techniques (e.g., feature transformation, dimensionality reduction, testing various time windows) to improve the accuracy of assessment models for accessing everyday behavioral patterns in order to predict their general social functioning. This work also proposes three scenarios for understanding performance in different potential clinical settings. It suggests researchers think about the subsequent design and development of possible intervention or support tools. This detailed investigation of how the models might be applied in real-world clinical settings, and understanding the extent to which they may or may not work in various contexts.
3.6 Discussion

3.6.2 Implications for HCI and Design

We believe our approach opens up new forms of social functioning assessment and intervention in the wild to improve peoples’ lives. Specifically, we imagine a “sensing core” (that can accurately assess social function in a passive low burden manner) integrated into existing and future health care systems. For populations living serious mental health issues as discussed in this chapter, our approach could be a game-changer: many people living with schizophrenia for example live on the edges of our society (like many of the people in our study), many do not have access to good healthcare services or even a stable place to live – fewer than 40% of people living with schizophrenia have contact with mental health services for treatment [92]. In the future, we imagine phone-based mental health assistants will combine passive assessment and social functioning intervention (e.g., akin to cognitive processing therapy) to inform and help people. There are many different design options to consider in the future from standalone systems to ones integrated with care support systems.

We also believe automated tools can be designed to be integrated into domain expert systems using our methods. For example, utilizing PCA of sensor data, clinicians can understand the patterns associated with higher and lower social functions of populations of clients in specific locations, and can be provided with more accurate interventions and suggestions for their target communities by the system – for example vulnerable communities, such as those with mental health issues, aging in place, homeless, etc. We imagine that various providers could integrate core social functioning sensing and intervention technology into their systems – for example, the core could trigger intervention and an early warning sign to clinicians when an outpatient has a significant deterioration of social functions. Furthermore patients,
3.6 Discussion

or other at need groups, such as elders, can also benefit from accurate assessment, self-reflection and intervention tools. Using our approach, individuals with social functioning impairment are able to learn from the behavioral patterns of groups with higher social functioning, and potentially avoid patterns of groups with lower social functioning. For example self-tracking of social functioning requires advances in how such data could be presented to the user in an informative and validated manner. Such challenges are at the core of HCI design. How do you present data and hints in an effective manner? How do you take into account that no design or model fits all?. Our approach advances how future systems can be designed. Many mobile intervention systems (e.g., [296, 219]) include predefined conversational therapies, however, they need considerable user input to provide suggestions. Our methods can be integrated into intervening HCI designs [223, 4, 282] to automatically find out personalized strategies according to individuals’ current status and improve social functions.

3.6.3 Implications for Sensing Populations at Risk

The participants in our study live extremely challenging lives and we were aware that many had not used a smartphone before and importantly did not fully understand the types of data being collected. Because of this we spent a considerable amount of time and effort educating participants about smartphones and the data being collected. Our team included computer scientists, psychiatrists, clinical psychologists, IRB specialists and the head of psychiatric research at Zucker Hillside Hospital, which specializes in the treatment of schizophrenia. Collectively the team had considerable experience working with vulnerable populations and mobile technology. For example,


3.7 Conclusion

Our mobile UI (e.g., self-reports, turning on/off sensor streams, etc) was designed by a clinical psychologist who has specialized in the development of mobile technology for communities with severe mental illness for two decades. It is important that participants fully understand the implications and risks associated with consenting to a mental health study in the wild. As part of the education and consent process, we screened out people incapable of understanding mobile technology, the data being collected and risks associated with participation.

3.7 Conclusion

We recognize the limitations of our work. (1) We only have 160 social functioning reports to train our models on. All patients live in a large dense city and the models may not generalize to other locations, such as, patients living in rural communities. (2) Smartphone sensing, though captures a variety of personal behaviors, is still limited to capture the fine-grained items in the questionnaire. For example, according to Table 3.5, the recreation sub-scale has the worst performance. It might be because mobile phone sensors are not able to capture detailed items in recreation such as “swimming”, “gardening”, etc. In the future, we could study additional sensing modalities (e.g., wearables, in-house sensors) for more accurate predictions. Due to privacy requirement we only collect the metadata and do not look at the exact content in social media, SMS and audios). In the future, we could study privacy-preserving NLP technologies to boost performance. (3) Some of the analysis and discussions imply some assumptions that may not be generalizable. For example, most of the employees in our study have full-time jobs and are working at regular hours as cashiers, construction workers, etc. Consequently, our interpretation of employment
sub-scale is based on regular job hours. However, the results and interpretations may not be generalizable when applied to another dataset, where people could work on non-traditional hours. Researchers should pay attention to this when they generalize the results in this chapter. (4) The performance from LOSO reveals that it is still challenging to accurately predict a new patient’s social functioning without any historical SFS records. This is inevitable at this stage - we have made the best effort to enroll the participants and collect the year-long data, but the dataset is still small. (5) The real-world impact of the prediction errors is still unclear. For example, from the data science perspective, MAE of 6.6 for recreation is 13.7% error and the lower the MAE, the better the prediction. However, from a strictly clinical perspective, the impact is not yet clear due to the lack of substantial research on the subject. That is why we hope we could open new directions to be precisely studied in future work.
Chapter 4


4.1 Introduction

Auditory Verbal Hallucinations (AVHs) are a sensory experience ranging from auditory imagery and vivid thoughts to fully developed hallucinations of hearing sounds and voices in the absence of any speaker [81, 174]. Although people with serious mental illness are more likely to experience AVHs (e.g., 60–80% of people diagnosed with
4.1 Introduction

schizophrenia [9]), AVH is also experienced by people with a variety of mental health diagnoses (e.g., anxiety, depression, functional impairment [157, 114]) as well as by people who have no identifiable psychiatric or neurological diseases [195, 311, 79]. It has been observed that individuals living with AVH require varying levels of treatment and needs and their clinical status can change throughout the course of their lives [174]. As a result, AVH represents a complex set of experiences that occur in a variety of forms across a wide range of people. Given this, there is a need to better understand the impact and severity of AVH across different population groups from those diagnosed with serious mental illness to those that have no such identifiable psychiatric impairment [81].

Retrospective self-report measures are commonly used in clinical settings to gain insights into the subjective experiences of auditory verbal hallucinations (AVH). However, these measures pose unique challenges. Patients with severe AVH symptoms often come from economically disadvantaged and socially isolated communities that lack adequate treatment resources [294, 230]. Additionally, some patients may be reluctant to seek treatment due to disagreement with the diagnosis or negative attitudes towards mental health services [206]. Those with less severe AVH may choose not to engage in traditional clinical settings due to stigma-related concerns [133, 119]. To address these limitations, recent clinical research projects have turned to the use of mobile phones and wearable technology to gather continuous behavioral data [20]. These studies have demonstrated connections between mobile sensing and behavioral markers of mental health [286, 240, 16, 338].

Despite the advantages offered by existing mobile sensing platforms, such as StudentLife [342], Aware [106], Beiwe [256], and others, in accurately monitoring pa-
4.1 Introduction

tients’ behavior in everyday life, there has been limited work on analyzing patients’ speech collected from mobile phones. The result is that current mobile sensing-based approaches are overlooking valuable information that could be obtained from everyday human speech in the naturalistic environment. Research suggests that speech patterns, such as incoherence [10], limited pitch variation [148], prolonged pause duration [277], and lexical diversity [364, 78, 69], can provide valuable insights into mental health. For example, Minor et al. [238] used an "electronic audio recorder" in real-world settings to better understand psychosis risk, demonstrating the feasibility of using audio recordings to assess real-world expressions of personality and functioning in schizotypy. However, the predictive value of this approach has not yet been determined. This chapter addresses the question of whether combining the power of speech in the wild with mobile sensing technology can improve our understanding of AVHs, and how much merit it can bring to the current mobile sensing system.

In this chapter, we report on the discriminative power of using speech and content (e.g., transcription) to assess the severity of AVH. We augment the StudentLife [342] mobile sensing app to solicit speech via a daily voice diary Ecological Momentary Assessment (EMA). Figure 4.1 shows the user interface of the daily voice diary. The voice EMA randomly prompts the user four times per day between the hours of 9 am and 9 pm. Participants can also open the app at any time to manually respond to the voice diary survey. The EMA allows participants to record a brief audio diary detailing how they are feeling (see Figure 4.1). The question that prompts the user is quite open in nature. Furthermore, we tell participants not to overthink how they present their thoughts: “It does not have to be perfect. In fact, we don’t want it to be. We want it to be real, and true to you. What would you like to share with the research
4.1 Introduction

team?”. In our study, we collected one month of mobile sensing data from N=384 participants. More importantly, we obtained a high volume of voice in the naturalistic environment: a total of 4809 voice diaries were collected, of which 3033 are longer than 30 seconds. While we collect and analyze all StudentLife sensing streams (e.g., phone usage, conversational interaction, mobility, activity, sleep), we focus on the power of human speech to assess and predict the severity of AVH. We use the Hamilton Program for Schizophrenia Voices Questionnaire (HPSVQ) [333] as ground truth that measures the severity of AVH. The HPSVQ is a 9-item questionnaire with a five-point Likert scale rated from 0 to 4. It has excellent test-retest reliability ($r = .84$) and internal consistency (Cronbach’s $\alpha = .94$). A total HPSVQ score above 26 indicates severe AVH. The HPSVQ scores are dichotomized to derive discrete categories as targets for classification. We design a specialized deep learning architecture to handle multimodal sensing streams (including audio, text and other behavioral sensing data) to predict AVH severity. We do this to best understand the discriminative power of voice. Specifically, we compare and contrast the performance of deep learning models built from (i) only mobile sensing data; (ii) only voice diaries (i.e., audio and transcribed text); and finally (iii) a fusion of mobile sensing data and voice diaries. We also consider both manually and automatically transcribed voice diaries in our modeling and analysis.

The contribution of the chapter is as follows. First, to the best of our knowledge, we are the first group to study how in the wild daily voice diaries can improve current mobile sensing platforms and predict AVH severity with good performance. We extend an existing and widely used mobile sensing platform for mental health with voice diaries and evaluate various deep neural network architectures step by step,
4.1 Introduction

beginning with Bi-direction Gated Recurrent Units (Bi-GRU) [65] and progressing to models with self-attentive embeddings [205], in order to understand the strengths and weaknesses of using each type of data (i.e., sensor data, audio and transcripts) and to find the best architecture that can fully utilize such rich and diverse multi-modal data for assessing AVH severity. Specifically, we propose a deep neural network architecture with self-attention that achieves a weighted f-1 score of 0.84 combining sensor data, audio data of the speech and professionally generated transcripts, and a macro f-1 score of 0.81 when manually generated transcripts using an automated speech recognition transcription method. A deep neural network model based solely on mobile sensing achieves a weighted f-1 score of 0.65, while a model based solely on voice diary (i.e., audio + text) achieves a weighted f-1 score of 0.78. Our results

Figure 4.1: User interface of the daily voice diary in study app. (A) Welcoming page (B) Recording page.
4.2 Related Work

indicate that periodic voice diaries combined with deep learning alone have sufficient power for assessing a complex condition such as AVH in the wild.

The structure for the rest of the chapter is as follows. We describe the related work in Section 4.2. In Section 4.3, we discuss our study and data collection. We describe how we preprocess our data in Section 4.4 and Section 4.5. We present our deep learning architecture and performance results in Section 4.6. We investigate the interpretation in Section 4.7. We discuss the performance of using a fully automatic pipeline using Automated Speech Recognition, as well as the implication of our work in Section 8. Finally, we offer some concluding remarks in Section 4.8.

4.2 Related Work

Mobile technologies have the advantage of passively collecting context-based in-situ data, which can be utilized to make various inferences regarding user behavior. The widespread use of smartphones, equipped with a range of embedded sensors such as accelerometers, GPS, microphones, and cameras, provides an opportunity to gain deep and continuous insights into the behavior of individuals diagnosed with mental health illnesses. Clinical research projects have recently embraced mobile phones and wearable sensing technology to collect continuous behavioral data. Studies have found correlations between behavioral sensor data obtained from smartphones and various mental health conditions, including anxiety [286, 38], depression [288, 347, 244, 263, 350], bipolar disorder [134, 284], and serious mental health conditions such as schizophrenia [25, 346, 26, 349]. These findings demonstrate the potential of mobile technologies for facilitating mental health diagnosis and management.

Much of the prior research in the field of mental health and speech primarily oper-
4.2 Related Work

ates in controlled laboratory settings, such as recording individuals with schizophrenia reading emotionally neutral text [277]. These studies have demonstrated that linguistic content (e.g., lexical diversity, patterns of word usage, semantic coherence, verbal fluency) [364, 78, 69] and acoustic parameters (e.g., intonation, pitch, jitter, energy [76]) can both provide insights into mental health. Some early attempts have been made to collect speech-related features via mobile sensing. In one of the initial studies, Muaremi et al [245] employed phone call statistics, social signals extracted from phone conversations, and acoustic features of the voice to predict bipolar states in a cohort of 12 individuals who suffered from bipolar disorder. The CenceMe app [237] implemented an always-on voice activity detector and higher-level conversation classifier on the phone to passively determine how much conversation a user is around. Early work by Lu [213] used passive speech analysis to model stress on an Android phone using prosody, pitch, volume, and intonation. Recently, Verily found [252] pairwise correlations between four basic audio parameters (average pause duration, duration of self-recorded audio) and transcribed text (speaking rate, sentiment score) from a 12-week depression research. These works are limited in scope and provide only a limited understanding of speech-related behaviors. To date, there has been no systematic investigation into the full predictive power of speech in natural settings and how it can be leveraged through mobile sensing technology.

Within the field of neural science, there are proponents who assert that auditory verbal hallucinations (AVH) are indicative of an auditory dysfunction [231]. However, the majority of scholars associate this form of hallucination with a speech disorder [116, 155]. It has been discovered that the dysfunction of brain regions responsible for speech production may serve as a fundamental mechanism for the occurrence of
4.2 Related Work

AVH in individuals with schizophrenia [316]. Empirical data suggests the presence of electrophysiological abnormalities [199] in the cortices responsible for auditory and speech perception, specifically in Wernicke’s area, which is associated with speech perception, and Broca’s area, which is associated with speech expression [150, 285]. Studies on auditory perception in individuals with schizophrenia spectrum disorders have revealed impairments in tone matching and pitch discrimination performance for non-verbal sounds [91]. Neuroimaging studies have also found revealed that in individuals with AVH, neural activation is lateralized towards the language-related areas of the right hemisphere [312]. However, in the majority of individuals, language production is primarily associated with activation in the left hemisphere [191]. The linguistic abilities of individuals diagnosed with schizophrenia have been found to be compromised in several ways, including decreased embedding, shorter phrase length, and increased grammatical errors [113, 243]. Individuals who experience clinical AVH tend to utilize a reduced number of determiners and prepositions, employ shorter utterances, and exhibit a greater prevalence of negative content [70].

The rapid advancement in speech processing technology enables the use of virtual assistants such as Google Assistant, Amazon’s Alexa, and Apple’s Siri [197]. These virtual assistants rely on deep learning algorithms and always-on speech processors to recognize hot words (e.g., Ok Google) and initiate interaction with the user. However, this method of collection is not practical for mental health studies due to the inability to differentiate between consented and non-consented speech. Instead, we consider Ecological Momentary Assessment (EMA), an alternative approach for collecting real-time data (including human speech) on AVH. Previous studies using EMA have demonstrated correlations between AVH and factors such as time of day [34],
4.3 Study and Dataset

Anxiety [147], emotional state [84], and cardiac autonomic control [181]. A mobile application that employs EMA to collect AVH-related ground truth from participants has also been developed [308]. Given the potential of vocal and lexical features of speech as predictors of AVH, a useful next step involves examining predictors of AVH using these attributes.

4.3 Study and Dataset

4.3.1 Procedures and Participants

The study was approved by the Institutional Review Board (IRB) of the University of Washington and Dartmouth College. Participants were recruited remotely over the Internet or through community-based strategies. Participants had to be at least 18 years old, English speakers residing in the United States, and have AVH at least once a week. It was also required that they have an Android phone and a data plan. Online recruiting was carried out with the help of Google Ads. Google displayed our ads to those whose search history matched pre-defined keywords. We chose keywords based on a study of the literature [74], consultation with researchers on mental illness, and a survey of blogs written by people living with AVH. The keywords contained both clinical terminology (e.g., schizophrenia, bipolar disorder, hearing voices) and non-clinical explanations of AVH (e.g., talking to ghosts, going crazy). If participants clicked on the Google advertisements, they were sent to our

---

1The National Institute of Mental Health’s Research Domain Criteria (RDoC) [77] is a research framework that guides AVH research on a continuum [111]. Our research, informed by the RDoC framework, focuses on the phenomenology of AVH to elucidate the distinctions between individuals who experience severe AVH symptoms in the context of a clinical disorder and those who encounter AVH without the need for clinical intervention. Therefore, the screening questionnaire is designed to exclude non-AVH individuals, as they fall outside the scope of this study.
4.3 Study and Dataset

recruitment website. The website included detailed infographics and videos that de-
scribe the projects in depth. On our website, participants were able to verify their
phone number, answer screening questions, agree to the study, complete the baseline
evaluation, and download the study Android app. Community-based recruiting
were through flyers and referrals. Our research staff prescreened the participants over
phone and scheduled in-person study visits. During the visit, research staff directed
participants to the study’s website and assisted them with the same procedures as
those who were recruited online.

A total of 384 individuals with AVH (305 recruited online; 79 recruited in the
local community) from 41 US states participated in the study and completed data
collection. All participants were required to carry a smartphone and respond to EMAs
for 30 days. Only if a participant adhered to the 30-day data collection period is he or
she regarded to have completed the data collection (those who dropped out during the
study are not included in the 384 participants). Participants could directly contact
research personnel to ask questions or get technical help. The data were firstly stored
locally on the phone and transferred to secure research servers when the phone detects
a connection to the Internet. The study app ceased delivering data to the study team
after 30 days. All participants were compensated $125 for participating. Table 4.1
shows the demographics of the participants.

4.3.2 Ground Truth

We administered the Hamilton Program for Schizophrenia Voices Questionnaire (HPSVQ)
[333] once at the start of the study period (i.e., during the enrollment). HPSVQ is
a 9-term self-report measure that received psychometric support [333] for assessing
4.3 Study and Dataset

Table 4.1: Demographics of the participants.

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>192</td>
<td>50.0%</td>
</tr>
<tr>
<td>Male</td>
<td>176</td>
<td>45.8%</td>
</tr>
<tr>
<td>Transgender: MTF</td>
<td>7</td>
<td>1.8%</td>
</tr>
<tr>
<td>Transgender: FTM</td>
<td>5</td>
<td>1.3%</td>
</tr>
<tr>
<td>Other:</td>
<td>4</td>
<td>1.0%</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>238</td>
<td>62.0%</td>
</tr>
<tr>
<td>Black or African American</td>
<td>80</td>
<td>20.8%</td>
</tr>
<tr>
<td>Asian</td>
<td>7</td>
<td>1.8%</td>
</tr>
<tr>
<td>American Indian / Alaska Native</td>
<td>6</td>
<td>1.6%</td>
</tr>
<tr>
<td>More than one race</td>
<td>49</td>
<td>12.8%</td>
</tr>
<tr>
<td>Missing / Declined</td>
<td>4</td>
<td>1.0%</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic / Latino</td>
<td>58</td>
<td>15.1%</td>
</tr>
<tr>
<td>Not Hispanic / Latino</td>
<td>324</td>
<td>84.4%</td>
</tr>
<tr>
<td>Missing / Declined</td>
<td>2</td>
<td>0.5%</td>
</tr>
</tbody>
</table>

AVH severity. The nine 5-point rating scale components quantify AVH features such as frequency, negative content, loudness, duration, interference with life, distress, influence on self-appraisal, and command compliance. A higher score indicates more severe auditory hallucinations. HPSVQ has shown high test-retest reliability and internal consistency. Furthermore, it has exhibited good concurrent validity [333] when compared to the widely used interviewer-rated Psychotic Symptoms Rating Scales (PSYRATS) [136]. HPSVQ indicates a total score of 0 to 13 for minimal/mild, 14 to 25 for moderate, and 26 and above for severe levels. The ground truth of prediction was derived from these categories.

According to the HPSVQ, 27.4% of individuals in our study exhibited severe AVH (score > 25). Those with scores <= 25, encompassing minimal, mild, or moderate
4.3 Study and Dataset

AVH, are classified as having non-severe AVH. Severe AVH is associated with depression, functional impairment, and schizophrenia spectrum disorders (SSDs) [157] and often requires clinical intervention. Identifying factors that distinguish severe AVH holds clinical significance, as it may contribute to the development of targeted treatments for individuals with AVH requiring intervention, as well as preventive strategies for subclinical populations so they do not progress in the psychotic trajectory. Consequently, we opt for a meaningful threshold suggested by the HPSVQ for our severe/non-severe AVH binary classifier’s ground truth, rather than employing a threshold from our dataset that yields balanced 50-50 labels.

4.3.3 Dataset

The Android study application is an upgraded version of a system used and validated in prior research [342, 24, 348, 239, 329, 349] that has been tailored to the demands of this study. The study application captures signals from a variety of embedded sensors (e.g., GPS, screen lock/unlock, light, microphone, accelerometer) to infer participant behavior (e.g., mobility, phone usage, physical activity, etc). We further developed a dashboard for research assistants to monitor user compliance.

We collected 30-day passive sensing data from 384 participants. We asked participants to optionally record a voice diary to share what they were experiencing or anything else they want to share with the research team. The voice EMA randomly prompts the user four times per day between the hours of 9 am and 9 pm. Participants can also open the app at any time to manually respond to the voice diary survey. Participants can record a voice diary of up to 3 minutes in length. The voice diaries are stored on the phone and uploaded to our secure server backend.
4.3 Study and Dataset

(a) Number of voice diaries recorded at different time of the day

(b) Duration of voice diaries recorded

Figure 4.2: We received 4809 audio diaries in all. (a) The majority of the voice diaries were recorded between the hours of 9 a.m. and 9 p.m., when the voice EMA randomly prompts. Others were recorded before or after the typical EMA time. (b) Many voice diaries are less than 30 seconds and contain simply random noise or a few words that are insufficient for analysis. There are also a significant number of audio diaries ranging in length from 2.5 minutes to 3 minutes (the maximum length allowed by the application).
4.3 Study and Dataset

Although the voice diary is optional and no extra incentives were offered for it, most participants submitted their recordings. We received a total of 4809 voice diaries, the distribution of which is depicted in Fig.4.2. There are a number of voice diaries that are less than 30 seconds long and contain only random noise or a few words that are insufficient for analysis and are thus eliminated from the analysis in this study. As such short recordings seldom contain interpretable language, we restrict this collection to recordings lasting 30 seconds or more, leaving 3033 records from 229 users. These 3033 voice diaries were professionally transcribed. The analysis in this chapter is based on the 229 users who provided complete 30-day mobile sensing data as well as at least one transcribed voice diary. Among the 229 users, 31.9% exhibited severe AVH, a proportion slightly higher than the 27.4% ratio observed among the total recruited participants, as discussed in Section 4.3.2. In this chapter, we refer to "audio/speech data" as the audio signals from the recordings and "text data" as the transcribed text from the voice diaries. In addition, we used our in-house automatic voice recognition system to create another batch of transcriptions. We also include an analysis in which we compare the results of employing manually and automatically generated transcripts (Section 4.7.2).

Fig.4.3 illustrates the word cloud of the audio diaries. Intriguingly, the word cloud plots of people with non-severe and severe AVH are similar; they both notably include keywords such as 'voice' and 'hear,' indicating that we need to develop a more comprehensive approach of AVH prediction than only examining the frequency of such words.
4.4 Data Preprocessing

This data set includes mobile sensing data, audio data from voice diaries, and transcribed text. In this section, we demonstrate how data preprocessing is used to transform unstructured data collected in the wild into well-formed data sets that can be used for further analytics.

Deep learning algorithms, including RNNs, have recently demonstrated exceptional performance without the need for handcrafted features when dealing with sequential data such as text streams, audio snippets, video clips, and time series data [298]. However, because deep machine learning is a black-box technology, the majority of individuals are unable to link the results to an explanation that is human-
4.4 Data Preprocessing

comprehensible. Such interpretable descriptions are essential for establishing trust between patients and clinicians in the context of practical health.

In our study, we employ both interpretable feature engineering and data preprocessing for deep learning. We perform feature engineering in order to (1) determine the benchmark performance of AVH severity prediction with traditional machine learning models and (2) provide interpretations for representations learned by deep learning models. In addition, we perform data preprocessing for deep learning so that multi-modal time series data can be properly fed to advanced deep models.

4.4.1 Interpretable Feature Engineering for Traditional Machine Learning

Acoustic Features

We extract acoustic features from speech signals using the open-source toolkit openSMILE (open-source Speech and Music Interpretation by Large-space Extraction) [104, 103]. OpenSMILE is extensively used for automatic emotion recognition in affective computing. We compute acoustic features included in the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [102], a state-of-the-art standard acoustic parameter set for various areas of automatic voice analysis, such as paralinguistic or clinical speech analysis.

First, 18 low-level descriptors (LLDs) are calculated on each frame (a short fragment that contains a few thousands of samples) of audio signals, including:

6 Frequency related parameters

Pitch: logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz.
4.4 Data Preprocessing

*Jitter:* deviations in individual consecutive F0 period lengths.

*Formant 1, 2, and 3 frequency:* center frequency of first, second, and third formant.

*Formant 1:* bandwidth of first formant.

### 3 Energy/Amplitude related parameters

*Shimmer:* difference of the peak amplitudes of consecutive F0 periods.

*Loudness:* estimate of perceived signal intensity from an auditory spectrum.

*Harmonics-to-noise ratio (HNR):* relation of energy in harmonic components to energy in noise-like components.

### 9 Spectral parameters

*Alpha Ratio:* ratio of the summed energy from 50-1000 Hz and 1-5 kHz.

*Hammarberg Index:* ratio of the strongest energy peak in the 0-2 kHz region to the strongest peak in the 2-5 kHz region.

*Spectral Slope 0-500 Hz and 500-1500 Hz:* linear regression slope of the logarithmic power spectrum within the two given bands.

*Formant 1, 2, and 3 relative energy:* the ratio of the energy of the spectral harmonic peak at the first, second, and third formant’s center frequency to the energy of the spectral peak at F0.

*Harmonic difference H1-H2:* ratio of energy of the first F0 harmonic (H1) to the energy of the second F0 harmonic (H2).

*Harmonic difference H1-A3:* ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3).
4.4 Data Preprocessing

Next, based on the 18 LLDs on each frame from an utterance, we generate the 62 acoustic features of high-level statistics functions (HSFs) that will be used in this chapter, including:

36 parameters of arithmetic mean and coefficient of variation computed through smoothing all LLDs over time with a symmetric moving average filter 3 frames long.

16 additional functionals based on loudness and pitch over voiced regions representing the 20th, 50th, and 80th percentile, the range of 20th to 80th percentile, and the mean and standard deviation of the slope of rising/falling signal parts.

4 additional parameters over all unvoiced segments: the arithmetic means of the Alpha Ratio, the Hammarberg Index, and the spectral slopes from 0-500 Hz and 500-1500 Hz.

6 temporal features, including the number of loudness peaks per second, the mean length and the standard deviation of continuously voiced regions the mean length and the standard deviation of unvoiced regions (approximating pauses), and the number of continuous voiced regions per second (pseudo syllable rate).

Linguistic Features

Linguistic Inquiry and Word Count (LIWC) [265, 321] is a computerized text analysis software used to quantify language in written or spoken communication. LIWC is commonly used to examine language data and can be applied to the study of language
4.4 Data Preprocessing

as an indication of psychological, social, and health-related outcomes. We adopt pre-defined categories in LIWC that have been used in existing pieces of literature that examine mental health from languages [190, 289]. Specifically, we use five categories of, in total 59 linguistic features: (1) **summary language variables**, (2) **basic linguistic style**, (3) **cognitive measures**, (4) **concerns**, (5) **language formality**.

**Summary variables** include 4 narrative evaluation: *analytical thinking* [266], *clout* [176], *authenticity* [251], and *emotional tone* [67] and *Word per sentence*. Then, we consider **basic linguistic style** from two aspects: percent of functional words and time orientations. We choose them because the structure of the word and tense are an expression of personality and inner thoughts [267]. Next, we export the **cognitive features**: cognitive processes and perception processes, which directly mirror human perception and thought. In addition, we consider 3 categories of **psychological concerns**: personal concerns, social concerns, and biological concerns as a participant’s concerns may reflect their mental states. Lastly, we assess the **formality of language**, which includes informal speech and all punctuation. These linguistic measures, which are outlined in Table 2, are extracted using the computerized text analysis software LIWC2015 [265, 321].

**Sensor Data Features.**

We compute features as described in Chapter 1 (Section 1.2.3), except that we do not compute location semantics in this study. Features (except for the sleep features) are computed on a daily basis and broken down into four epochs of the day: *morning* (6am-12pm), *afternoon* (12pm-6pm), *evening* (6pm-12am) and *night* (12am-6am), that allow us to model people’s behaviors during different parts of the day as recom-
4.4 Data Preprocessing

Table 4.2: Linguistic features

<table>
<thead>
<tr>
<th>category</th>
<th>feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>summary language variables</td>
<td>analytical thinking, clout, authenticity, emotional tone, word per sentence (wps)</td>
</tr>
<tr>
<td>basic linguistic style</td>
<td>1st person singular, 1st person plural, 2nd person, 3rd person singular, 3rd person plural, impersonal pronouns, articles, prepositions, auxiliary verbs, adverbs, verbs, adjectives, comparisons, interrogatives, numbers</td>
</tr>
<tr>
<td>time orientations</td>
<td>past orientation, present orientation, future orientation</td>
</tr>
<tr>
<td>cognitive measures</td>
<td>insight, causation, discrepancy, tentative, certainty, differentiation</td>
</tr>
<tr>
<td>perceptual processes</td>
<td>see, hear, feel</td>
</tr>
<tr>
<td>personal concerns</td>
<td>work, leisure, home, money, religion, death</td>
</tr>
<tr>
<td>social concerns</td>
<td>family, friends, female references, male references</td>
</tr>
<tr>
<td>biological concerns</td>
<td>body, health, sexual, ingestion</td>
</tr>
<tr>
<td>language formality</td>
<td>swear words, netspeak words</td>
</tr>
<tr>
<td>punctuation</td>
<td>periods, commas, colons, semicolons, question marks, exclamation marks, dashes, quotation marks, apostrophes, parentheses, other irregular marks</td>
</tr>
</tbody>
</table>

mended by multiple existing research [345, 349]. As a result, we came up with a total of 103 features. Following that, we compute the arithmetic mean of each feature over the 30 days to represent a participant’s overall behavior during the research. Days with less than 15 hours of collected data (approximately 9.2% of all days) are considered to have too much missing data and are omitted from the calculation of mean values [346].
4.4 Data Preprocessing

4.4.2 Data Preprocessing for Deep Learning

Acoustic Signals from Voice Diaries

We must first transform the acoustic signal into a series of numbers before we can use it in our models. Our voice diary signals are sampled at 44100Hz. In other words, a single one-second clip would contain 44100 samples, which are too numerous to be processed in their raw state. In fact, any sound frame can be represented by combining signals of various frequencies. And the spectrum - the combination of frequencies that comprise a signal - can be used for further analysis in audio processing using deep learning. Additionally, humans do not perceive frequencies linearly. Variations at lower frequencies are perceived more readily than those at higher frequencies. Similar to frequency, humans perceive loudness logarithmically rather than linearly. Consequently, when dealing with frequencies and amplitudes in our data, we should utilize a logarithmic scale by means of the Mel Scale and the Decibel Scale. Mel Spectrogram is the outcome of these two scale adjustments. We compute the Mel Spectrogram using Librosa, a python-based audio signal processing library. Fig 4.4 shows the Mel Spectrogram computed from one of our voice diaries. It plots frequency along the y-axis and time along the x-axis.

We define a frame as a sequence of 8192 values, which corresponds approximately to a 200ms segment in a 44100Hz diary clip. The maximum length of a participant’s diary is 3 minutes, resulting in a maximum length T of 180s * 44100Hz / 8192 = 968 in our computed output. We compute 128 Mel scales for each frame (the default setting of Librosa, which is commonly used in audio signal processing). As a result, the entire output (which will be used as the input for deep models) is a \((T = 968) \times (M = 128)\) output.
4.4 Data Preprocessing

Figure 4.4: Mel Spectrogram computed from one of our voice diaries.

= 128) matrix, and the colors in Fig. 4.4 represent the values. The outputs of diaries with a duration of less than 3 minutes are zero-padded so that the padded values can be identified and bypassed in RNN.

Transcripts from Voice Diaries

Before we can feed the transcripts into our models, we must first convert them to a set of vectors. We use the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model [88] to extract contextual embeddings from the text. BERT is trained on the Wikipedia corpus in order to solve different natural language processing tasks (e.g., named entity recognition). We first tokenize the words in the sentences based on BERT’s vocabulary. We then extract a contextualized embedding for each of the tokens. BERT automatically splits the original word into smaller subwords and characters (i.e., tokens). If the input has multiple sentences, BERT uses the special token [SEP] to differentiate them. It also inserts a [CLS] token at the start of the text to indicate the beginning. Each token is then embedded into a 768-dimensional space.
4.4 Data Preprocessing

producing a single 768-length vector. As a result, for each transcript text file, we create a sequence of length $T$, where $T$ is the number of tokens separated by BERT, and each specific position $t$ in a such sequence is a vector of length 768. All created sequences are 0 padded afterward to match the length of the longest text in the data set, as we must keep the output the same shape for each transcript to feed them into the deep models.

Sensor Data from Smart Phones

Mobile sensing data is commonly used as sequences in two ways: (1) low-level raw sensor signals sampled every second, and (2) high-level aggregated/inferred behavioral variables on an epoch/daily basis. The first method is typically utilized for real-time applications like activity recognition. The second is usually applied in long-term research. We use the second strategy in this study and create a time sequence with a length of $T=30$ (days). In such a sequence, each position $t$ is a vector containing 103 features generated in Section 4.4.1. In deep learning, we keep the entire sequence instead of computing the arithmetic mean of each feature throughout the 30 days so that the deep model can learn knowledge from the entire time series. Deep neural networks frequently employ normalization procedures to boost convergence and generalization. As a result, for each feature, we compute two normalized (within-person and between-person) values to capture both within-person and between-person differences during the study. Finally, the dimension of the sensor data input is $T \times M$, where $T = 30$ denotes the number of days in the research and $M = 2 \times 103 = 206$ denotes the number of normalized features computed from sensor signals. All missing values are replaced with -999 so that they can be recognized and skipped in deep
4.5 Predicting Models and Results

In the following sections of this chapter, we refer to the pre-processed acoustic signals, transcripted text, and sensor data that we plan to use as inputs to our deep models as "audio data," "text data," and "sensor data" for simplicity.

4.5 Predicting Models and Results

In this section, we evaluate various deep neural network architectures, from simple to complex, in order to comprehend the predictive power of each type of data (sensor data, audio data, and text) and to identify the best architecture capable of utilizing such rich and diverse multi-modal data for assessing AVH severity. The binary labels we used for training and tried to predict are based on the HPSVQ participants completed during the enrollment, as described in Section 4.3.2. That says, the prediction is made once per participant over the whole duration of the study, which is 30 days.

To evaluate the models, we selected 60%, 20% and 20% of the independent users, respectively, as the training, validation, and test sets. Data from a single participant appear in only one of the sets (leave-subjects-out). We are aware that each user has multiple sequences of processed audio data and transcripts (depending on the

---

Identify severe AVH using the 1-month dataset holds diagnostic value. Per the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [13], the co-occurrence of AVH symptoms during a 1-month period, along with another “characteristic sympto” such as delusions or disorganized speech, is indicative of a diagnosis of schizophrenia. Furthermore, it should be noted that the HPSVQ is not intended to measure short-term changes, as evidenced by its consistently high test-retest reliability when administered on a weekly basis [180]. In light of the characteristics of AVH, researchers scholars commonly assess the changes in the phenomenological characteristics of AVH through longer-term follow-up, such as 6 months [180] or even up to 1 year [43].
number of voice diaries uploaded), but only one sequence of 30-day sensor data (from Section 3). We are not using all of the audio data and transcripts to train the model because (1) some participants may upload significantly more voice diaries than others, resulting in an unbalanced training set (with a large amount of data from a single participant), and (2) short voice diaries may not contain sufficient information for data analysis. Therefore, we select the 10 voice diaries with the longest audio length for each participant, and then select the 5 with the highest word count in the transcript. This strategy may not be ideal, but it should be satisfactory for choosing a set of most informative voice diaries from each participant, given that we do not intend to infer the occurrence of AVH from individual audio diaries. At each training epoch for users in the training set, one of the user’s selected voice diaries is chosen to train the model. For each user in the validating or testing set, we simply select a voice diary at random from those that have been selected. The model is trained with the Adam optimizer [182] and tuned for optimal performance on the validation set. The testing set performance is reported.

4.5.1 Baseline: Using Traditional Machine Learning Approaches

Before moving on to deep learning approaches (which typically outperform regular machine learning methods on sequential data), we would like to establish a baseline with classic machine learning methods. We train an XGBoost [61] classifier with all of the interpretable features specified in Section 4.1. XGBoost is a highly efficient and flexible gradient-boosting library. We conduct a grid search on specified hyperparameters to optimize the performance of the model formed from the training set.

\footnote{It is assumed that the lengthy voice diaries selected during the month are the most representative for inferring the severity of AVH, while shorter ones may introduce noise to the prediction.}
4.5 Predicting Models and Results

on the validation set. Using the selected super-parameters, we next train a model on
the combined training and validation data and evaluate it on the testing set. Our
XGBoost model achieves a macro f1-score of 0.68 and a weighted f1-score of 0.72. We
will further discuss the important features in Section 5.5.4.

4.5.2 Uni-modal Data

In this subsection, we use the pre-processed audio data, transcripts, and sensor data
from Section 4.4.2 as the input for deep learning models to evaluate the performance
of predicting AVH severity from each data type. A masking layer [218] is used im-
mediately after the input to inform the model that certain timesteps in the input are
missing (e.g., the 0 padding) and should therefore be skipped when processing the
data.

Bidirectional Gated Recurrent Unit (BiGRU)

We first predict the AVH using a Bidirectional Gated Recurrent Unit (BiGRU) net-
work [65]. If there are H hidden units in the GRU, the bidirectional GRU would
output 2 * H values. The model is constructed as detailed in Fig.4.5. The orange
rectangle represents a vector of length M at a particular time for a sequence shaped
T * M. The bidirectional GRU layer produces an output (the red rectangle) with the
shape (T, 2 * H), where H is the number of hidden units in the GRU. Thus, we create
a simple vector representation of the sequence data. The representation generated
by the bidirectional GRU layer is then fed into fully connected neural networks and
dropout layers.
4.5 Predicting Models and Results

Figure 4.5: Illustration of Bidirectional Gated Recurrent Unit network.

**Self-attentive Embedding**

We augment the bidirectional GRU models with self-attention embedding [205], a widely employed technique for sentence embedding in natural language processing. This attention mechanism enables a recurrent model to focus on distinct portions of a lengthy input stream. The model is constructed in accordance with Fig.5.5. We anticipate that self-attentive embedding will improve performance when text data is used as input because it has been utilized successfully in text processing. We have yet to determine whether or not it improves performance when audio or sensor data is used as input.

Instead of relying solely on the final hidden state of the RNN, self-attentive embedding necessitates all hidden states at each step of the RNN. Assuming that the hidden unit number for each unidirectional GRU is $H$, we will obtain an output of $B = (h_1, h_2, ... h_T)$ from BiGRU, where $h_t$ represents the concatenation of the hidden states from the forward and backward GRU at position $t$. The attention mechanism encodes $B$ into an embedding by choosing a linear combination of the $T$ BiGRU hid-
4.5 Predicting Models and Results

Figure 4.6: Illustration of self-attentive BiGRU network.

Hidden vectors in B. The linear combination weights are defined in the attention vector $\alpha$, which is computed as $\alpha = \text{softmax}(w_2 \ast \text{tanh}(w_1 \ast B^T))$. The $w_1$ and $w_2$ are learned from the data themselves. Note that $w_2$ is trained with an initial input of zero because we have no prior knowledge of the attention vector. The softmax ensures that the sum of all computed weights is 1. Then, weighted sums are computed by taking the inner product of the attention vectors $a$ and $B$. The representation produced by self-attentive BiGRU is then used as the input for subsequent fully connected neural networks and dropout layers to predict the severity of AVH.

Predicting AVH using Uni-modal Data

Table 4.3 displays the f-1 scores when predicting the severity of AVH using uni-modal data (sensor, audio, or text) and various deep neural networks. As anticipated, our experiments demonstrate that self-attentive models, which were originally utilized in NLP areas, dramatically improve the performance of text data-based models. The
4.5 Predicting Models and Results

text-based self-attentive model achieves a close to 0.7 weighted f-1 score. It marginally improves the audio data a little bit (in terms of macro f-1). This indicates that despite the fact that the mel spectrogram of raw audio data still contains a great deal of prosody-semantic information, it is more difficult to learn a good attentive embedding from the spectrogram than from the transcribed text. In terms of sensor data, the self-attentive model performs even worse than the BiGRU model. This suggests that focusing on different areas of the 30-day study period will not improve the AVH prediction, which is to be expected given the nature of rolling enrollment. Instead, we could simply apply RNN without attentive embedding.

Table 4.3: Predictive performance on AVH using uni-model data.

<table>
<thead>
<tr>
<th></th>
<th>bigru-sensor</th>
<th>bigru-audio</th>
<th>bigru-text</th>
<th>self-att-sensor</th>
<th>self-att-audio</th>
<th>self-att-text</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1-score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>macro</td>
<td>0.59</td>
<td>0.53</td>
<td>0.52</td>
<td>0.46</td>
<td>0.56</td>
<td>0.68</td>
</tr>
<tr>
<td>weighted</td>
<td>0.65</td>
<td>0.63</td>
<td>0.6</td>
<td>0.58</td>
<td>0.58</td>
<td>0.7</td>
</tr>
</tbody>
</table>

4.5.3 Multi-modal Data

In this subsection, we combine the optimal deep learning architecture for each type of uni-modal data described in Section 4.5.2 by concatenating the representations from the sensor, audio, and text neural networks prior to feeding them into the final fully connected neural networks and dropout layers (Fig. 4.7a). In addition, we hypothesize that para-linguistic features (e.g., emotions, speed, energy, tone, etc.), which may be involved in the representation of audio (orange rectangle in Fig. 4.7b, may influence the attentive embedding in text. Consequently, in Fig. 4.7b, we use the representation of the audio as prior knowledge about the attention vector for text and learn \( w_2 \) (one of the parameters for generating the attention vector, see Fig. 5.5)
4.6 Interpretation

starting from this representation rather than a zero input.

Table 4.4 depicts the F-1 scores for predicting the severity of AVH using multi-modal data and the two architectures depicted in Fig. 4.7. Architecture (b) that uses the audio sequence representation as a starting point for learning the text attentive embedding parameter outperforms architecture (a). The audio + text model achieves a weighted f-1 score of 0.78, indicating the discriminatory power of using the wild voice diaries to determine the severity of AVH. In addition, the addition of passive sensor data from mobile devices eventually results in a better weighted f-1 score of 0.84. Table 4.5 and Table 4.6 compare the confusion matrices between the "audio+text" model and the "combine-all" model, using the architecture (b). Based on the analysis of the confusion matrices, it appears that the inclusion of mobile sensing data has the potential to enhance precision, but does not appear to have a significant impact on recall within our dataset. In our best model, the false positive rate, which refers to the probability of an individual being erroneously diagnosed with severe AVH, is minimal, specifically 6%.

Table 4.4: Predictive performance on AVH using multi-model data.

<table>
<thead>
<tr>
<th></th>
<th><code>audio+text (a)</code></th>
<th><code>combined all (a)</code></th>
<th><code>audio+text (b)</code></th>
<th><code>combined all (b)</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>macro</td>
<td>0.73</td>
<td>0.77</td>
<td>0.75</td>
<td>0.81</td>
</tr>
<tr>
<td>weighted</td>
<td>0.75</td>
<td>0.81</td>
<td>0.78</td>
<td>0.84</td>
</tr>
</tbody>
</table>

4.6 Interpretation

In this section, we study the feature interpretation for AVH prediction in two ways: the importance of handcrafted features in the XGBoost classifier and the interpretation of deep learning-learned features.
4.6 Interpretation

Figure 4.7: Illustration of deep neural network using multi-modal data (audio, text and sensor data).

4.6.1 Important Interpretable Features

SHAP (SHapley Additive exPlanations) is a game-theoretic way to explaining the output of any machine learning model [214]. SHAP is one of the most used post-hoc
4.6 Interpretation

Table 4.5: Confusion matrix: using architecture (Fig. 4.7b) and the "audio + text" model (weighted f-1 score = 0.78)

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Prediction</th>
<th>0</th>
<th>1</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>True Neg.</td>
<td>26</td>
<td>5</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>False Pos.</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>False Neg.</td>
<td>5</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>True Pos.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>31</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.6: Confusion matrix: using architecture (Fig. 4.7b) and the "combined-all" model (weighted f-1 score = 0.84)

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Prediction</th>
<th>0</th>
<th>1</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>True Neg.</td>
<td>29</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>False Pos.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>False Neg.</td>
<td>5</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>True Pos.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>34</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

explainability methods for computing feature attributions. To understand how a single feature affects the AVH prediction, we plot the SHAP values of top-15 important features for every sample in the dataset in Fig 4.8. The y-axis, from top to bottom, ranks the features in order of importance. The x-axis refers to the actual SHAP values. The horizontal location of a point shows the feature's impact on the model's prediction for a given sample (x > 0: more likely to be severe AVH; x < 0: more likely
4.6 Interpretation

to be non-severe AVH). The color of each point represents the value of the feature, with red indicating a high value and blue indicating a low value. This visualization provides a comprehensive overview of the contributions of each feature to the AVH prediction.

Figure 4.8: SHAP value summary plot of the XGBoost model trained on handcrafted features

From the summary plot presented in Fig. 4.8, it can be observed that the following features have a significant impact on the prediction of a person with severe AVH:

[Audio diary text features] lower occurrence of words related to “hearing things”; higher occurrence of words related to “death”, 3rd person plural, and “seeing things”; having higher clout (power in talking).

[Audio diary acoustic features] larger number of continuous voiced regions per second, smaller mean of the deviations in individual consecutive F0 period lengths,
4.6 Interpretation

higher SD on the ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3), higher mean Hammarberg index (the ratio of the strongest energy peaks in the 0–2 kHz vs 2–5 kHz regions) of unvoiced regions, higher loudness, lower first formant (F1) frequency, shorter length of continuously voiced regions.

[Mobile sensing features] higher deviation in illuminance during night; fewer call-ins during night.

The results of the study provide new insights into AVH, some of which align with established clinical knowledge and hypotheses. For example, words pertaining to “death” may serve as an indicator of the presence of major depressive disorder [22]. “Seeing things” may indicate the presence of visual hallucinations, which typically co-occur in association with other hallucinations such as AVH [353]. The larger number of continuous voiced regions per second and shorter length of continuously voiced regions are consistent with the tendency of clinical AVH patients to employ shorter utterances [70]. Other acoustic features may suggest basic impairments in auditory processing [330] or voice-related distress [306]. The deviation in illuminance during the night and fewer call-ins could indicate social withdrawal [156]. All of the aforementioned factors are associated with AVH. However, some of the results may appear counter-intuitive, such as the lower occurrence of words related to “hearing things” among patients with severe AVH. This may be due to the fact that patients with more severe symptoms often have lower insight, and therefore do not describe their experiences within a clear frame of “hearing things”. Some may still be lacking in literature support (e.g., higher clout). Despite these limitations, the important features identified in this study demonstrate the feasibility of using multi-modal data
4.6 Interpretation

for assessing the severity of AVH. However, it is important to note that the feature importance is based on a baseline model and further validation is needed through additional research to avoid false discoveries. We plan to delve deeper into interpreting the high-level representations automatically learned through deep learning.

4.6.2 Interpretation of Auto-Learned Features from Deep Learning

It is crucial that we open the “black box” of a deep neural network and explain why it makes decisions in health and wellness applications in order to enhance patient and clinician trust. In this part, we propose a post hoc analysis for clinical interpretation of the learned representations from voice, transcript, and sensor data obtained from our best model (reference Fig. 4.7b). We export the values in the hidden units in the yellow layer (256 hidden units that represent audio input), green layer (256 hidden units that represent text input), red layer (256 hidden units that represent sensor data input), and the fully connected layer (256 hidden-units that represent multi-modal data input) before the last dropout layer. We label each value in the learned representations with the type of data source and its index in the representations; for instance, audio-i denotes the i-th value in the representation learned from audio sequence (yellow rectangle in Fig. 4.7b) and text-j denotes the j-th value in the representation learned from text sequence (green rectangle in Fig. 4.7b). We employ ANOVA to examine the auto-learned features across groups with non-severe AVH and severe AVH. Based on the effect size (partial eta squared) from ANOVA, the most discriminative auto-learned features of each category of data are chosen. Then, we analyze the relationship between these selected auto-learned features and
4.6 Interpretation

interpretable hand-crafted features generated in Section 4.4.1.

Representations Learned from Audio Data

Fig 4.9 shows the distribution of the top-3 most discriminative auto-learned features from audio data (represented by the green rectangle in Fig. 4.7b) among the non-severe (label 0) and severe (label 1) AVH groups. People with severe AVH are more likely to have a higher value in audio-19, audio-135 and audio-60. The effect sizes are all around 0.06, indicating a medium effect.

The table 4.7 lists the top-3 hand-crafted features with the strongest correlations (i.e., having the highest absolute values of correlations) to the selected auto-learned audio features. For a detailed explanation of each hand-crafted feature, refer to Section 4.4.1. Our results suggest that the network has learned multiple representations associated with lower pitch and loudness, as well as steeper slopes of both rising and falling portions of loudness. These representations may indicate negative symptoms or distress, respectively, and correspond to severe AVH.

These findings align with previous research in psychology and neuroscience. For instance, individuals with schizophrenia spectrum disorders often experience deficits in auditory perception, such as pitch-based tone matching [318]. Additionally, these individuals may have event-related potential deficits and neuroanatomical abnormalities in the auditory cortex [168]. Such deficits in basic auditory processing and sensation can impact higher-level cognition and have been linked to hallucinations [167]. Studies have also shown that individuals with auditory hallucinations perform worse on pitch discrimination tasks compared to those without AVH [232].
4.6 Interpretation

Figure 4.9: The distribution of the most discriminative representations from audio data among the non-severe (label 0) and severe (label 1) AVH group. The effect size is indicated by partial eta squared.

(a) audio-19, $\eta^2 = 0.07$

(b) audio-135, $\eta^2 = 0.06$

(c) audio-60, $\eta^2 = 0.06$
4.6 Interpretation

Table 4.7: Top-3 representations automatically learned from audio data of voice diaries that identify patients with severe AVH, as well as the top-3 handcrafted acoustic features with the strongest correlation to the representations.

<table>
<thead>
<tr>
<th>Representations</th>
<th>feature_1</th>
<th>feature_2</th>
<th>feature_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>audio-19</td>
<td>(+0.26) Mean Hammarberg index over voiced regions</td>
<td>(-0.26) Mean Spectral Slope 0-500 Hz over voiced regions</td>
<td>(-0.24) Mean Pitch</td>
</tr>
<tr>
<td>audio-135</td>
<td>(-0.26) Mean Spectral Slope 0-500 Hz over voiced regions</td>
<td>(+0.23) Standard deviation of the slope of rising signal parts of Loudness</td>
<td>(+0.22) Mean slope of rising signal parts of Loudness</td>
</tr>
<tr>
<td>audio-60</td>
<td>(+0.23) Standard deviation of the slope of falling signal parts of Pitch</td>
<td>(+0.22) Mean slope of falling signal parts of Pitch</td>
<td>(-0.18) Mean Loudness</td>
</tr>
</tbody>
</table>

Representations Learned from Text Data

Fig. 4.10 illustrates the distribution of the top 3 most discriminative auto-learned features from text data (represented by the green rectangle in Fig. 4.7b) between non-severe (label 0) and severe (label 1) AVH groups. Individuals with severe AVH tend to exhibit higher values for features text-39, text-93, and text-13. The effect sizes suggest a large or close-to-large effect. The network has learned high-level representations that are associated with text data that contains more past tense, fewer present tense expressions, a higher emphasis on personal concerns related to money and work, and fewer references to cognitive and perceptual processes, which may indicate the presence of severe AVH (refer to Table 4.8).

The focus on personal concerns related to money and work may suggest stressors that could exacerbate symptoms [307]. There have been studies on the relationship between past focus and mental illness such as schizophrenia and mood disorder, and
4.6 Interpretation

have led to mixed results, with some research showing an increased use [87] and others a decreased use [19] of past tense in written materials focusing on schizophrenia. It’s worth noting that these studies relied on written materials, such as online essays and social media content, while our analysis is based on transcripts from oral audio diaries. The findings regarding a lower emphasis on cognitive and perceptual processes are consistent with the argument by Javitt [167] that sensory deficits can negatively impact cognitive processes.

Table 4.8: Top-3 representations automatically learned from text data from voice diaries that differentiate patients with severe AVH the most effectively, as well as the top-3 created linguistic features with the strongest correlation to the representations.

<table>
<thead>
<tr>
<th>Representations</th>
<th>feature_1</th>
<th>feature_2</th>
<th>feature_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>text-39</td>
<td>(+0.37) Past tense</td>
<td>(+0.30) Personal concerns: Money</td>
<td>(-0.29) Present tense</td>
</tr>
<tr>
<td>text-93</td>
<td>(-0.31) Perceptual processes: Hear</td>
<td>(+0.27) Personal concerns: Money</td>
<td>(-0.25) Perceptual processes: All</td>
</tr>
<tr>
<td>text-13</td>
<td>(+0.27) Personal concerns: Work</td>
<td>(+0.27) Personal concerns: Money</td>
<td>(-0.27) Perceptual processes: All</td>
</tr>
</tbody>
</table>

Representations Learned from Mobile Sensing Data

Fig 4.11 shows the distribution of the top-3 representations auto-learned representations from mobile sensing data (red rectangle in Fig. 4.7b), obtained from smartphones, that effectively distinguish patients with severe AVH. Patients with severe AVH tend to have lower values for sensor-4, sensor-183, and sensor-181. The effect sizes, as measured by partial eta squared, suggest a medium effect.

Table 4.9 lists the top-3 hand-crafted features with the strongest correlation to the auto-learned representations from mobile sensing data. It is important to note that the coefficient listed in the table represents the association between the hand-crafted
4.6 Interpretation

Figure 4.10: The distribution of the most discriminative transcript representations between the non-severe (label 0) and severe (label 1) AVH groups, effect size indicated by partial eta squared.
4.6 Interpretation

Figure 4.11: The distribution of the most discriminative sensing data representations between the non-severe (label 0) and severe (label 1) AVH groups.

(a) sensor-4, $\eta^2 = 0.08$

(b) sensor-183, $\eta^2 = 0.08$

(c) sensor-181, $\eta^2 = 0.07$

features and the auto-learned representations. Given that severe AVH is correlated with lower values of the representations (Figure 4.11), readers should consider the in-
4.6 Interpretation

verse coefficients when examining the relationships between the hand-crafted features and severe AVH.

Table 4.9: Top-3 hand-crafted features with the strongest correlation to the auto-learned representations from mobile sensing data.

<table>
<thead>
<tr>
<th>Representations</th>
<th>feature_1</th>
<th>feature_2</th>
<th>feature_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>sensor-4</td>
<td>(-0.27) Duration of conversations nearby</td>
<td>(-0.27) Sleep: wake up time</td>
<td>(-0.27) Number of conversations nearby</td>
</tr>
<tr>
<td></td>
<td>(night, 0am-6am)</td>
<td></td>
<td>(0am-6am)</td>
</tr>
<tr>
<td>sensor-183</td>
<td>(+0.48) Duration of out-going phone calls</td>
<td>(+0.45) Duration of out-going phone calls</td>
<td>(+0.45) Number of in-coming phone calls</td>
</tr>
<tr>
<td></td>
<td>(whole day)</td>
<td>(morning, 6am-12pm)</td>
<td>(whole day)</td>
</tr>
<tr>
<td>sensor-181</td>
<td>(+0.35) Duration of phone usage (evening,</td>
<td>(+0.32) Duration of in-coming phone calls</td>
<td>(+0.32) Duration of phone usage (whole day)</td>
</tr>
<tr>
<td></td>
<td>6pm-0am)</td>
<td>(whole day)</td>
<td></td>
</tr>
</tbody>
</table>

Sensor-4 represents night-time conversations and wake-up time, both of which may indicate insomnia, a known association with auditory hallucinations [304]. Sensor-183 pertains to telephone calls, while Sensor-181 focuses on mobile device usage. In general, the findings suggest that sleeping in and waking up late, reduced incoming and outgoing phone calls, and decreased phone usage could indicate severe AVH, potentially reflecting social isolation [156], which has been proposed to have a mechanistic relationship with psychotic symptoms.

Figure 4.12: Attention vector of a segment of an audio diary.
4.7 Discussion

4.6.3 Learned Attention Vector in Sentence Embedding

The use of self-attention in text enables an interpretation of the learned sentence embedding: we can generate heat maps of the weight vectors associated with the text, allowing us to visualize which portions of the audio diary are considered when predicting AVH. In Figure 4.12, we present the attention vector for a selected segment of the audio diary. The darker areas on the heat map indicate the clauses or phrases that carry more weight in the attention vector. The model appears to capture significant factors in the audio diary that are related to the severity of AVH. These factors include following voices ("maybe it’s that I get control"), interfering voices ("but the thing is that it starts to come on strong"), loud/distressing voices (feeling like you want a gallon scream at the top), and feeling bad of self ("I’m so frustrated, it’s a very confusing illness"), all of which are encompassed in HPSVQ.

4.7 Discussion

4.7.1 Using Multimodal Streams and Deep Learning to Predict AVH

The purpose of our research was to identify the best deep learning architecture for predicting AVH using multimodal stream data. To accomplish this, we conducted experiments using various deep learning approaches.

In contrast to traditional machine learning algorithms, deep learning models have the advantage of being able to handle time series data more effectively. Traditional algorithms require manual extraction of a large number of statistical features from
4.7 Discussion

each time series, leading to loss of information and a high number of correlated features. Moreover, there is no way to assign different time periods in the series different weights, because it is difficult to determine which time periods are most relevant for predicting outcomes. Our experiments showed that deep learning models outperformed XGBoost, confirming their suitability for this type of problem.

Traditional machine learning algorithms may struggle with time series data. Researchers must compute multiple hand-crafted or aggregated statistical features on each time series data (e.g., average, percentile, maximum, minimum, etc.), resulting in information loss and an explosion of highly correlated features. In addition, there is no way to assign different periods in the time series different weights because we do not know which time period is most relevant for predicting a particular outcome. Our experiment also shows that deep learning outperforms the XGBoost.

The results presented in Table 4.3 demonstrate that the text-based model is more effective compared to those relying on speech and sensor data. This highlights the significance of participant’s words (i.e. content) as better indicators of AVH, compared to the acoustic parameters in speech and daily behavior recorded by mobile phone sensors. This finding is noteworthy, particularly in light of previous studies that have predominantly focused on predicting symptoms of schizophrenia through passively sensed behavior [329, 346, 6]. However, predicting the severity of AVH can prove to be a more complex task, as it is a symptom that can occur not just in individuals with serious psychotic disorders like schizophrenia, but also in healthy individuals. In fact, as many as 75% of individuals who do not meet the criteria for a psychotic disorder have reported experiencing AVH [175]. Moreover, although AVH is a hallmark symptom of schizophrenia, not all patients with schizophrenia experience AVH
4.7 Discussion

[62]. Some neuroscience studies have shown that compared to schizophrenia patients without AVH (non-AVH), those with AVH exhibit distinct intrinsic connectivity patterns in cortico-subcortical circuits [281] and interhemispheric circuits [60]. However, the underlying mechanisms by which AVH arise spontaneously from intrinsic brain activity remain elusive. Our results imply that predicting AVH based solely on passively sensed behavior may not be sufficient, and demonstrate the efficacy of utilizing voice diary sequences.

4.7.2 Fully Automatic Pipeline (Using Automated Speech Recognition)

So far, in our investigations, we have utilized text that has been manually transcribed. This method enables us to grasp the potential of voice diary transcriptions. Nevertheless, manual transcription services are not feasible due to their high cost and manual nature. Thus, we are now further exploring the use of Automated Speech Recognition (ASR) in a fully automated system.

With the aim of preserving patient confidentiality, we have chosen to utilize an in-house ASR system instead of relying on audio-to-text services from third-party providers. This in-house system [365] is based on Baidu’s Deep Speech 2 architecture [8] and was implemented using PyTorch [261]. The system was trained with speech corpora obtained from the Linguistic Data Consortium ⁴, a consortium that includes universities, libraries, corporations, and government research laboratories.

The results of our AVH prediction performance using a fully automated system with ASR instead of manual transcription services can be seen in Table 4.10. Although

⁴https://www.ldc.upenn.edu/
4.7 Discussion

it is expected that the accuracy of the prediction may decrease due to errors in the
ASR transcription process, our results show that even with sensor data, audio, and
transcripts obtained through this automated approach, we still obtained a macro f-
1 score of 0.76 and a weighted f-1 score of 0.80. These findings suggest that it is
possible for the entire AVH prediction process to be carried out with minimal human
involvement in the future.

Table 4.10: Predictive performance on AVH using multi-model data (with text data
from ASR). See Fig. 4.7 for the definition of structure (a) and (b).

<table>
<thead>
<tr>
<th>f1-score</th>
<th>audio+text (a)</th>
<th>combined all (a)</th>
<th>audio+text (b)</th>
<th>combined all (b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>macro</td>
<td>0.64</td>
<td>0.76</td>
<td>0.72</td>
<td>0.76</td>
</tr>
<tr>
<td>weighted</td>
<td>0.71</td>
<td>0.80</td>
<td>0.78</td>
<td>0.80</td>
</tr>
</tbody>
</table>

4.7.3 Implications

Despite the fact that researchers in the ubiquitous computing community (many of
whom partnered with clinical researchers) have investigated mobile sensing-based
technology for a number of years [20], it has not yet been widely adopted in the
clinical treatment process. This chapter presents an alternative method to enhance
existing mobile sensing-based applications by utilizing a simple periodic voice diary
to assess the severity of AVH in the wild.

AVH is the perceptual experience of hearing sounds or voices. This type of audi-
tory hallucination may cause considerable distress and disability. 5-28% of the general
population experiences AVH [173, 81, 325], However, it is challenging for medical pro-
fessionals to assess the severity of AVH. Not all patients wish to or are able to contact
physicians due to a variety of factors (e.g., stigma, discrimination. lack of access to
services). Fewer than 40% of people with schizophrenia receive treatment from men-
4.8 Conclusion

tal health professionals [92], let alone those with AVH, which can be experienced by individuals with no diagnosed psychiatric disorders. Theoretically, the proposed automated systems can be reached remotely and wouldn’t require as much from the individual relative to going to in-person treatment, which is more ecologically valid and scalable.

4.8 Conclusion

The treatment needs and clinical status of individuals with AVH can vary and change over time [174]. Our study highlights the usefulness and feasibility of using voice diary EMA to assess the severity of AVH. Table 4.3 demonstrates that the complete model, which integrates audio, text from voice diaries, and mobile sensor data, achieved the highest level of performance. However, it is noteworthy that acceptable results can still be obtained by solely utilizing audio and manually transcribed text from voice diaries. Moreover, the automated process presents the possibility of developing an AVH predictive tool with minimal human intervention. The methodology presents new opportunities for the deployment of health-centric mobile sensing applications, offering a potentially more energy-efficient and privacy-conscious alternative. Furthermore, the successful online recruitment of participants in our study shows their willingness to use EMA and smartphone technology for self-tracking of AVH. Voice diary platforms possess the potential to assume a pivotal role in conducting comprehensive mental health evaluations in the future.
Chapter 5

First-Gen Lens: Assessing Mental Health of First-Generation Students across Their First Year at College Using Mobile Sensing

5.1 Introduction

The transition from high school through the first year at college is particularly difficult for students [40, 44, 68]. While many students quickly adapt and excel during their first year navigating various challenges on campus, other first-years experience increased stress, isolation, loneliness, anxiety and depression. In a survey of more than 150,000 students, 9.5% of the first-year students report feeling frequently depressed, whereas 34.6% report being overwhelmed by academic pressures and other demands.
5.1 Introduction

[94]. It is important that students know how to address challenges and risks as they start their first year to grow as healthy young adults and succeed both socially and academically. Among all first-year students entering college, “first-generation” college students are one group that experience increased risk. These are students who come from families with no history of college degrees. Therefore, first-generation students have no family history of how to deal with the risks and demands of challenging academic environments, how to fit in, and how to cope with the various challenges they will encounter [86, 278, 224]. The risk of attrition in the first year among first-generation students is 71% higher than that for non-first-generation students [165].

Compared to their non-first-generation peers, first-generation students are more likely to face challenges that jeopardize their abilities to adjust to college life and achieve academic success [161].

Researchers have identified various risk factors associated with first-generation students, including, physical and mental health, family support in decision-making [371], social support and socioeconomic status (SES) [45, 371]. The higher education literature uses at-risk as a term for students who are poorly equipped to deal with challenges at college [161]. In [158], the authors discuss how first-generation students do not receive the same level of social support from their families and friends as their non-first-generation counterparts. The authors also report that first-generation students are more likely to experience depressive symptoms in comparison to non-first-generation students. First-generation students are also known to spend less time socializing with peers and interacting with teachers [322] which limits their network. Prior research also finds that students from lower socioeconomic backgrounds struggle with fitting in at college. Getting help, assessing options and planning for college are
5.1 Introduction

all novel challenges for them, making the transition to college even more difficult [75]. Given the enormous challenges facing first-generation students, it is not surprising to learn that they experience higher attrition rates [235, 32] and poorer mental health [169] in comparison to non-first-generation students.

While there is a wealth of literature demonstrating the need for educators and researchers to pay attention to the mental health of first-generation college students, there is no work that tracks and forecasts the mental health of this at-risk group. Beyond surveys, self-reports and anecdotal demographic descriptions concerning the stresses and strains of first-generation college life, we use a passive sensing tool and predictive analytics to offer insights into the mental health (e.g., depression, anxiety) of first-generation students. Specifically, we design a low-cost mental health prediction system for students at risk that aims to address the following research questions:

- Researchers have reported various risk factors associated with first-generation students. We aim to replicate these findings using mobile sensing technology.

  **(Q1)** Are the first-generation students in our cohort at-risk? Do such risk factors have an association with their mental health during their first year at college? And, if so, how? More importantly, which of these risk factors might be most associated with mental health?

- Given these risk factors, we hypothesize that first-generation students exhibit behavioral patterns distinct from non-first-generation students in the first year.

  **(Q2)** What are the key behavioral differences between first-generation and non-first-generation students across each term as they progress through their first year? And how are these behaviors associated with risk factors?
5.1 Introduction

- We hypothesize that some first-generation students learn new coping skills to overcome barriers and challenges [86, 274] during their first year. As a result, they are less at-risk, enabling them to better transition to college and manage their mental well-being effectively like many non-first-generation students.

(Q3) What behaviors do first-generation students exhibit that are associated with better mental health? In particular, are the behaviors of these students that cope better similar to the behaviors exhibited by non-first-generation students?

- The insights and outcomes from addressing the prior questions offer an opportunity to model and predict mental health from sensor data. One modeling challenge is overcoming the in-balance and possible bias between the minority population (i.e., 27 first-generation students) and the majority population (i.e., 153 non-first-generation students) in the modeling cohort.

(Q4) Can we accurately predict the mental health of the first-generation students using deep learning by taking into account important distinguishing behavioral factors of first-generation students?

In recent years, passive sensing using mobile smartphone technology has enabled users to assess daily behaviors without burdening the user. For example, the StudentLife study [342] established the first link between passively sensed activities and mental health outcomes for college students. However, a common shortcoming of existing research is that it focuses on general enrolled students without considering the distinctive characteristics of certain groups, such as first-generation students. In addition, prior studies of college students have not considered the full first year of college, which is a critical period of transition for all students but particularly for
5.1 Introduction

first-generation students. As a result, the field is unable to gain an in-depth understanding of first-generation students’ behavior and mental health across their full first year at college. In this chapter, we study N=180 first-year students using mobile phone sensing throughout their first year. Notably, 15% of the 180 first-year students (N=27) are first-generation students. To control for selection bias as much as possible, the study was advertised to all first-year students during the enrollment period at Dartmouth College irrespective of whether they were first-generation students or not; that is, we avoid explicitly selecting only first-generation students. Even with such selection criteria in place, our First-Gen study population is proportionally representative of the number of first-generation students admitted to Dartmouth College on an annual basis; in fact, we have 15% first-generation students in our cohort of students, which is slightly higher than the normal rate admitted to the university (which varies between 10-12 % annually).

To the best of our knowledge, this is the first study to investigate the first-year student experience using smartphone data across a full 12-month period, including all their academic terms and their academic breaks when students typically return home. The longitudinal nature of the First-Gen study offers an in-depth portrait of the first year of college life. It allows us to explore the behavioral patterns and differences at a level of detail not possible before. Furthermore, it presents an opportunity to study the predictive nature of time series sensing data and its relationship with the mental well-being of first-generation students. In addressing the research questions discussed above, we make the following contributions:

- We capture and quantify the high school years as first-year students enter Dartmouth College using a high school life survey. We assess the risk factors for
all students (N=180), taking into account various dimensions of the survey, including, socioeconomics, lifestyle, and social and support networks. We also use periodic Ecological Momentary Assessment (EMA) to collect self-reported mental health data from students. We find that first-generation students are at more risk based on their socioeconomic status, lifestyle and support network, but not in the area of sociability. Among these risk factors, however, lifestyle and sociability in high school (which are more about behavioral patterns than demographics) had a strong association with mental health. Lifestyle and sociability can be studied through passive-sensing behavioral data [140, 342]. This suggests that passive sensing may be beneficial in assessing mental health and that there might be some utility in exploring the use of passive sensing to understand the behavior of students.

- We use mobile sensing from students’ phones to capture the behavior of first-generation and non-first-generation students during their first year. There are significant differences in behavioral patterns between these two groups. For example, first-generation students spend more time in study areas and seldom visit the gym and Greek houses where campus-wide social events and parties usually occur. First-generation students also exhibit different behaviors than their non-first-generation counterparts. For example, they experience less regular sleep patterns and are more regular in the places they visit on campus.

- As each term of the academic year progresses, we capture changes in inferred behaviors associated with better mental health for first-generation students as they adapt to overcome initial risk factors. We find that such behaviors are not identical to the behaviors of the non-first-generation students; that is, it
5.1 Introduction

appears as though first-generation students have unique correlations between behavior captured by smartphone sensing and mental health. For example, a longer phone unlock duration at study places may indicate a deterioration of mental health for first-generation students, whereas spending less time at Greek houses and having a less regular lifestyle are strong indicators of poor mental health for non-first-generation students,

- Based on the observations discussed above, we design a novel deep learning model for mobile sensing time series data that pays attention to first-generation students when predicting their mental health. We propose a new deep model architecture that improves the overall F1-score by 0.07 – increasing it from the baseline of 0.63 to 0.70. More importantly, we discover that learning models based on generic architectures found in the existing literature are biased towards the majority (i.e., non-first-generation students) and perform poorly on the minority population (i.e, first-generation students). Our First-Gen deep learning architecture removes this bias and improves the F1-score for first-generation students from 0.58 to 0.71. We can boost the performance of our predictive model by capturing the important behavioral differences between the two student groups.

The structure of the chapter is as follows: We first discuss related work in Section 5.2, followed by visualization and initial analysis of first-year students in Section 5.3. Following this, we present the details of our First-Gen study in terms of study design, surveys, EMA design and feature extraction. As discussed above, we present our results in Section 5.5 and discuss how they support the four exploratory research questions that drive the study. In Section 5.6, we discuss our key research findings,
implications of the study on college administrations and limitations of the work. We finish with some concluding remarks in Section 5.7.

5.2 Related Work

The start of college life represents one of the most significant transitions in a young person's life. Fromme et al. [117] conducted an online survey of N=2025 students and investigated behavioral risks during the transition to college from high school life using six metrics; these are, alcohol use, drinking and driving, aggression, drug use, crime and the number of sexual partners. In general, the authors [117] find alcohol use, marijuana use and sexual relations increase due to newly acquired freedom by the students. In another study of N=1453 students, the authors [40] report that the performance on standardized tests (e.g., the SAT) predicts mental health in the transition period from high school to college. Some researches imply that higher peer ability (i.e., having more "advantaged" classmates who come from families with high levels of parental education and income [146]) has a detrimental impact on student outcomes such as educational aspirations. Self-concept theory, sometimes known as the "big fish, little pond" effect, is a classic psychological theory that explains how peer ability and the person’s place within the group’s ability distribution impacts individual learning [217]. As peer ability increases after high school, students’ sense of worth reduces due to the relative decline in their academic performance. This is akin to feeling like a “big fish in a small pond” and the sense by students that all “fish” is big at college. In addition, the first term of the first year plays an important role in the academic year. For example, many students are unable to reach their target GPA established at the beginning of their first term [183]. This phenomenon
5.3 First Year in a Nutshell

is partly due to students devoting a considerable amount of time during their first term to socializing and expending capital to establish their social network rather than dedicating time to studying and advancing their academic goals. As a result, students typically spend more time on academics during their second term onward [323].

There is considerable research on using mobile sensing to infer human behavior across various fields. In terms of passive sensing and mental health, many studies have emerged [139, 342, 248, 249, 351] reporting on depression [347, 288], anxiety [38, 270], stress [35, 250] and mood [203, 374]. The StudentLife study [342] first reported on how passively sensed behaviors from phones, such as conversational interaction, sleep and activity are associated with mental health outcomes and academic performance for 48 college students enrolled in one class over a single term. In another study of N=117 students living in a university’s dormitory, researchers [35] trained a Random Forest machine learning model based on the daily weather, personality of the students and smartphone sensing data to classify the stress level of students with an accuracy of 72%. Researchers also discovered growing evidence that mobility features relate to depression [53, 288, 287]. Finally, Xu et al. [366] conducted a study of N=188 undergraduate students and captured routine behaviors, as well as behavior pattern differences between depressive and non-depressive subgroups.

5.3 First Year in a Nutshell

In this section, we present an overview of the students included in the First-Gen study (N=180) across their entire first year at Dartmouth College. We present a set of time series visualizations of inferred student behaviors and self-reports representing their first year in a nutshell.
5.3 First Year in a Nutshell

5.3.1 College Life

Dartmouth College is a highly competitive academic institution located in a small college town in the northeast United States. Unlike universities in big cities, Dartmouth College is located on a self-contained campus. All first-year students are required to live on-campus, making on-campus mobile sensing meaningful because the campus is the very place where all student activities happen. Almost all the on-campus buildings are associated with a primary function, such as dorms, classrooms, library, gym, cafeteria, partying/social, etc. The number of multi-functional tall buildings on campus is negligible. Therefore, even without knowing the precise indoor location, the broader GPS signal indicates the primary type of area the students are in.

There are three major terms during the academic year at Dartmouth, each lasting ten weeks: the fall term (mid-September to November), the winter term (January to early March) and the spring term (late March to early June). There are three academic breaks during the year (viz. winter break, spring break and summer break) when undergraduates leave campus and typically return home, go on vacation, or take up internships in their hometown or away from home. All the terms are fast-paced compared to a 15-week semester system. Thus, students face an intense midterm period soon after the term begins. The major is not fixed for incoming students. During the first year, students need to seek out where their interests lie and the major they will work on in the years ahead. Students are recommended to select three courses each term. The largest on-campus social week of the year occurs in mid-May to bring the entire community together.
5.3 First Year in a Nutshell

5.3.2 First-Year Data

In what follows, we present trends in student behavior and mental well-being across each term and break for the incoming first-year students. Figure 5.1 shows the dynamics of mental health across the entire year for the cohort based on weekly self-reports and passively inferred behaviors (e.g., sleep duration, physical activity, etc.) obtained from students’ phones. Collectively, these time series plots provide an interesting glimpse into the complex lives of first-year students in the First-Gen study. Note, due to the rolling enrollment process, the plots in the first half of the fall term (Sep-Oct) are based on incomplete participants and may not be as reliable on the aggregate as the later months.

Figure 5.1a to 5.1d show the fluctuation of mental health over the year as measured by Ecological Momentary Assessments (EMAs). In the fall term, the term immediately after enrollment, we can see that depression and anxiety increase gradually, whereas self-esteem rapidly declines. We observe that first-generation students appear to have poorer mental health, particularly during the first term, probably due to the difficulty of adjusting to a highly competitive college environment. During the first winter break after the end of the first term, self-esteem improves dramatically and anxiety and stress fall to their lowest levels. When students are home, we can observe that their mental state rebounds back to their baseline as they decompress, sleep, and relax away from academic and other pressures (e.g., social, sports, financial, relationships). After returning to campus for the winter term, self-esteem starts to decline again while the depression, anxiety and stress increases. Same pattern also exists throughout the spring term. In particular, depression peaks in May, when self-esteem is at its lowest point. Throughout the entire year, we observe a pattern of
5.3 First Year in a Nutshell

![Figure 5.1: Mental health dynamics, sleep and physical activity over the first year. Plots are based on the mean value and 25/75 percentiles of each measurement among students, grouped by first and non-first generations. Participants are completely enrolled by mid-October. The plot indicates that the first half of the fall term (Sep-Oct) has insufficient participation.](image-url)
5.3 First Year in a Nutshell

mental health deteriorating as the term progresses and a bounce back during breaks. However, the degree of deterioration during the term is larger than the degree of recovery during holidays. As a result, mental health progressively deteriorates through the year. At the start of the longest break of the year (i.e., the 3-month summer vacation), depression, anxiety and stress dip to the same level observed during spring break – back to the baseline. These mental health indicators start rising before the start of the new term representing a U-curve in the figure. We also observe the same anticipatory mental health response before the students in our study return for their second year.

Figure 5.1e and 5.1f show how much students are sleeping and being physically active across the entire academic year from mobile phone sensing. During the fall term, sleep duration significantly drops at the beginning of the term, rises, and then falls to its lowest point. Physical activity also seems to be affected by the academic calendar as it declines during the midterm exam period and toward finals. During winter break, sleep duration increases sharply from its lowest point to its highest peak. We presume students sleep more because there is no pressure from classes, tests, exams and social demands back home. Physical activity also drops to its lowest point during the winter break. We are likely observing seasonal influences during the winter term and break, as shown by a decrease in mobility and physical activity of the students in our study. Sleep duration during the winter term declines as the classes begin. We also observe a drop in sleep duration during the midterm exam period. Sleep duration and physical activity increase towards spring break. In mid-May, physical activity peaks during the year’s biggest on-campus social event (i.e., the Green Key spring festival). This trend appears to be more pronounced among
5.4 First-Gen Methodology

non-first-generation students in our study cohort.

5.4 First-Gen Methodology

In this section, we discuss the details of the First-Gen study, mobile sensing and self-report system and data. Later, in Section 5.5 and Section 5.6, we discuss our results and insights, respectively.

5.4.1 Study Design

The First-Gen study recruits $N=180$ first-year students, among which $N=27$ are first-generation students, representing 15% of the total participants. Note that the percentage of first-generation students admitted to Dartmouth College varies between 10-12% per year on average. Therefore, we have a greater first-generation sample than the typical incoming class in any particular year, which adds power to the study. This study is approved by the Institutional Review Board (IRB) at Dartmouth College. Students were recruited and consented to participate at the beginning of their first year when joining Dartmouth College from high school, starting in September 2017. Students who agree to participate in the First-Gen study install our data collection application on their Android or Apple smartphone. The mobile sensing application collects two types of data: sensing data to capture users’ behavior and EMA to measure mental health. Participants were compensated for their weekly EMA responses at 10 dollars per week. The majority of our participants (67.8%, $N=122$) identify as female. In terms of race, 60% ($N=108$) are White, 23.9% ($N=43$) are Asians, 3.3% ($N=6$) are Black or African American, 2.8% ($N=5$) are American Indian/Alaska Na-
5.4 First-Gen Methodology

tive, 6.7% (N=12) belong to more than one race, and 3.3% (N=6) have not reported their race.

5.4.2 High School Survey

We use a high school survey to identify risks associated with students entering their fall term at Dartmouth College (Section 5.5.1). The survey, shown in Table 5.1, captures pre-college variables based on students’ self-reported family background and experience during the high school years. Survey questions are selected by the psychologists in our research team from the widely adopted CIRP Freshmen Survey (TFS) [283] and focus on four major risk factors often included in educational literature [161, 339, 356, 97, 96, 158, 371, 322]: (1) socioeconomic status (SES) [339, 356], (2) lifestyle [97, 96], (3) support [158, 371, 322] and (4) sociability [322]. The original TFS is intended for use with incoming first-year students before their first day of classes. It has been administered to over 15 million students at over 1,900 institutions for over 50 years. Importantly, it has provided valid and reliable data on incoming college students’ demographic characteristics, high school experiences, attitudes, behaviors, and expectations for college [303].

5.4.3 StudentLife: Behavioral Sensing App

App and System Design

We use the StudentLife mobile sensing app [342] for our First-Gen study. StudentLife has been used for several longitudinal studies in mental health across the United States [342, 344, 347, 346, 348, 349] and it allows us to sense students’ behavior using iOS and Android phones passively. We upgraded StudentLife to meet the demands of
5.4 First-Gen Methodology

Table 5.1: High School Survey based on the CIRP Freshmen Survey (TFS) [283]: Questions associated with risk factors.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Likert Scale Options (score 1-5)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Socioeconomic Status</strong></td>
<td>Lower socioeconomic class, Lower-middle socioeconomic class, Middle socioeconomic class, Middle-upper socioeconomic class, Upper socioeconomic class</td>
</tr>
<tr>
<td>I perceive myself as:</td>
<td></td>
</tr>
<tr>
<td><strong>Lifestyle</strong></td>
<td></td>
</tr>
<tr>
<td>During your time in HIGH SCHOOL:</td>
<td></td>
</tr>
<tr>
<td>How would you rate your physical well-being?</td>
<td>Very poor, Poor, Average, Good, Very good</td>
</tr>
<tr>
<td>How would you rate your mental well-being?</td>
<td>Very poor, Poor, Average, Good, Very good</td>
</tr>
<tr>
<td>How physically active were you?</td>
<td>Not at all active, Slight active, Somewhat active, Very active, Extremely active</td>
</tr>
<tr>
<td>Did you have healthy sleeping patterns?</td>
<td>Not at all, Slightly, Somewhat, Very, Extremely</td>
</tr>
<tr>
<td>Did you have healthy eating patterns?</td>
<td>Not at all, Slightly, Somewhat, Very, Extremely</td>
</tr>
<tr>
<td><strong>Support</strong></td>
<td></td>
</tr>
<tr>
<td>During HIGH SCHOOL, how often did you feel?:</td>
<td></td>
</tr>
<tr>
<td>That my family provided me with the support that helped me succeed.</td>
<td>Never, Seldom, Sometimes, Often, Almost always</td>
</tr>
<tr>
<td>That teachers provided me with feedback that helped me assess my progress in my classes.</td>
<td>Never, Seldom, Sometimes, Often, Almost always</td>
</tr>
<tr>
<td><strong>Sociability</strong></td>
<td></td>
</tr>
<tr>
<td>During HIGH SCHOOL, how often did you feel?:</td>
<td></td>
</tr>
<tr>
<td>Lonely or homesick</td>
<td>Never, Seldom, Sometimes, Often, Almost always</td>
</tr>
<tr>
<td>Isolated from school life</td>
<td>Never, Seldom, Sometimes, Often, Almost always</td>
</tr>
<tr>
<td>How satisfied were you with your social life?</td>
<td>Not at all, Slightly, Somewhat, Very, Extremely</td>
</tr>
</tbody>
</table>

longitudinal studies. We also developed a dashboard for research assistants to monitor user compliance. During the study, 85.2% of Android and 87.3% of iOS phones collect more than 23 hours of data per day. If we change the threshold to 19 hours, the
metric is 91.5% for Android phones and 93.8% for iOS phones. There is no significant difference in the data quality of the first-generation and non-first-generation students. Considering that few students turn off their phones during sleeping hours, we have one year of observations with high data coverage. At the end of their first year, 173 of the original 180 students completed the study, representing a retention rate of 96%, significantly higher than existing longitudinal mobile sensing studies listed in Table 5.2. Note that the other studies shown in Table 5.2 are for shorter duration periods. Those that provided monetary compensation, paid students at a similar or higher weekly rate than the First-Gen study (i.e., greater than $10 per week).

We conducted two pilot studies and solicited feedback before the main First-Gen study. The first pilot consists of a two-week focus group with 10 undergraduate students. Everybody is required to report the battery usage of our app daily, which is displayed on the system’s battery management page of the study dashboard. StudentLife consumes between 3% and 8% of the total energy. Almost all students claimed that they did not alter their charging practices due to installing StudentLife on their phones. In addition, students indicated that they did not anticipate issues with using the app for a longitudinal one-year study. For the second pilot, we recruited 45 students for four-week testing period of the app and system. Several software bugs were fixed during this period, and a 6% average battery usage was reported during the trial.

**Feature Extraction**

We generate several features from the collected mobile sensing data. A full list of behavioral features are shown in Table 5.3. In addition to the features described in
### 5.4 First-Gen Methodology

#### Table 5.2: Compliance in existing literature.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Target</th>
<th>Recruited participants</th>
<th>Duration</th>
<th>Compliance</th>
<th>compensation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xu et al. 2019 [366]</td>
<td>depression</td>
<td>188 + 267 college students</td>
<td>106 days</td>
<td>Phase 1: 73% and Phase 2: 79%</td>
<td>A Fitbit Flex 2 ($100) and cash to $205 based on compliance</td>
</tr>
<tr>
<td>Wang et al. 2018(a) [348]</td>
<td>personality</td>
<td>646 college students</td>
<td>2 weeks</td>
<td>159 participants (24.6%) gave more than 7 days with 19+ hours of sensing data</td>
<td>Students did self-tracking using apps as part of a course assignment</td>
</tr>
<tr>
<td>Wang et al. 2018(b) [347]</td>
<td>depression</td>
<td>83 undergrads. students</td>
<td>9 weeks</td>
<td>Not revealed in paper</td>
<td></td>
</tr>
<tr>
<td>Boukhechba, 2018 [37]</td>
<td>mental health</td>
<td>72 undergrads. students</td>
<td>2 weeks</td>
<td>EMA compliance 72%</td>
<td></td>
</tr>
<tr>
<td>Zhang, 2017 [374]</td>
<td>mood</td>
<td>42 college students</td>
<td>1 month</td>
<td>30 students (71.4%)</td>
<td>Not revealed in paper</td>
</tr>
<tr>
<td>Farhan, 2016 [105]</td>
<td>depression</td>
<td>79 college students</td>
<td>5 months</td>
<td>44.3% (GPS+PHQ9) 56.3% (Activity + PHQ9)</td>
<td>$15 Amazon gift card for every two weeks of active participation.</td>
</tr>
<tr>
<td>Huang, 2016 [162]</td>
<td>social anxiety</td>
<td>18 undergrads. students</td>
<td>10 days</td>
<td>Course credits and money (amount not revealed)</td>
<td></td>
</tr>
<tr>
<td>Wang, 2014 [342]</td>
<td>mental health, academic performance</td>
<td>60 college students</td>
<td>10 weeks</td>
<td>48 subjects (80%) completed</td>
<td>T-shirt. 10 Jawbone UPs and 10 Google Nexus 4 phones to top student collectors.</td>
</tr>
</tbody>
</table>

Chapter 1 (Section 1.2.3), we also compute several new features.

**Class Attendance.** We also compute class attendance based on students’ registered courses and classrooms provided by the registrar’s office. This strategy has been
5.4 First-Gen Methodology

Table 5.3: Features computed from StudentLife mobile sensing data. Daily and epoch-based sensing features are computed across meaningful daily epochs: night/ morning (12am-9am), day (9am-6pm), and evening (6pm-12am), allowing analysis of behavioral features and trends across different periods of the 24-hour clock).

<table>
<thead>
<tr>
<th>category</th>
<th>details</th>
</tr>
</thead>
<tbody>
<tr>
<td>physical activities</td>
<td>duration walking / running / cycling / in vehicle / sedentary</td>
</tr>
<tr>
<td>mobility and semantic locations</td>
<td>number of locations visited, distance travelled, max distance from the center of campus, duration at own dorms, other’s dorms, food area, gyms, study places, social places (Greek houses), class attendance.</td>
</tr>
<tr>
<td>phone usage</td>
<td>number of lock/unlocks &amp; unlocked duration at all places, study places, social places, own dorms, other’s dorms,</td>
</tr>
<tr>
<td>sleep patterns</td>
<td>sleep duration, sleep start time, sleep end time</td>
</tr>
<tr>
<td>regularity</td>
<td>location regularity (all categories, food and eating related, home related), circadian rhythm in physical activities, regularity of sleep duration, sleep start time, sleep end time over the past week.</td>
</tr>
</tbody>
</table>

validated during prior studies [344] at Dartmouth College.

**Regularity.** Additionally, we calculate the circadian rhythm in physical activities using the method described in [348]. We also compute location regularity. Previous studies [348, 53, 234] calculate location regularity based on whether participants were at the same GPS coordinates during the same period across different days. On the other hand, campus life has a distinct week-to-week rhythm, and buildings with similar primary functions are dispersed throughout campus. As a result, we compute the regularity of the locations using the Levenshtein distance [200] of the location categories across the same days on different weeks. We also compute the mean squares successive difference [336] (MSSD) of sleep duration and bedtime. MSSD measures the degree of autocorrelation and serves as an indicator of stability for things such as sleep. Regularity indexes are commonly explored in mental health sensing studies, and many of them report finding a significant association between the two factors [53, 253].
5.5 Analysis

5.4.4 Ecological Momentary Assessment (EMA): Self-reported Mental Health

We track the dynamics of student mental health using a built-in mobile EMA component integrated into the StudentLife app. Self-reports are randomly delivered to each student’s phone once every week. We landed on this as a period that could scale across a year and not be considered burdensome to users. We use a PHQ-4 [189] to evaluate depression and anxiety once per week. PHQ-4 asks questions about the frequency of experiencing “feeling nervous, anxious or on edge”, “not being able to stop or control worrying”, “feeling down, depressed or hopeless” and “little interest or pleasure in doing things” over the last two weeks respectively, all using a 4-point scale: 0: not at all, 1: several days, 2: more than half the days, 3: nearly every day. We collected 4942 EMAs from 180 students over the entire year of the First-Gen study.

5.5 Analysis

In this section, we discuss the results of the First-Gen study. Specifically, we address each of the driving research questions in turn (viz. Q1-Q4) as discussed earlier and shown in Figure 5.2.
5.5 Analysis

Figure 5.2: Flow of analysis of research questions addressed by the First-Gen study

5.5.1 Q1: How Do the Risk Factors Associated with First-generation Students Relate to Their Mental Health during Their First Year at College?

Researchers have proposed various risk factors associated with first-generation students. In what follows, we break down these known risk factors into four pre-college metrics, as discussed earlier: socioeconomic status (SES), lifestyle, sociability and support. We quantify these risk factors and show how they relate to mental health. To do this, we take the following approach: (1) we compare risk factor scores between first-generation and non-first-generation students using a t-test; (2) we compare the mental health (using PHQ-4) of first-generation and non-first-generation students for each term using mixed-effects models; and finally (3) we analyze how these pre-college risk factors are associated with mental health. This helps us better understand the challenges faced by first-generation students during their first year. We perform various hypothesis tests in the following sections. To make sure we identify actual relationships and not spurious ones, we create two stratified, randomly sampled, non-overlapping subsets of students (each with N=90 students) while maintaining a similar ratio of first-generation to non-first-generation students in both subsets – 77 non-first-
5.5 Analysis

generations and 13 first-generation in Subset A, and 76 non-first-generations and 14 first-generation in Subset B. Throughout the rest of the chapter, we refer to subset A using the symbol $\textcircled{A}$ and subset B using the symbol $\textcircled{B}$. We then test our findings in these subsets and verify whether they hold true for each subset. Furthermore, we use a two-stage Benjamini-Hochberg method (TSBH) \[29\] to calculate the false discovery rate (FDR), ensuring a rigorous approach to analysis.

Comparing Risk Factors between First-gens and Non-first-gens.

We discuss four risk factors collected from our high school survey. Table 5.4 shows the average score and standard deviation of each risk factor described in Table 5.1. A lower score on an item indicates that students in a particular group (viz. first-gens and non-first-gens) are at-risk for that factor. The scores are normally distributed ($p > 0.05$ in the Kolmogorov Smirnov normality test \[204\]). We compare the scores for each risk factor between first-generation and non-first-generation students using a t-test. Outlier points greater than (mean + 2 * std) or less than (mean - 2 * std) are eliminated before the test to avoid producing erroneous statistical results. The results show that the first-generation students have significantly lower socioeconomic status (SES), poorer physical and mental health (referred to as lifestyle in the table), and are less satisfied with support from their families and teachers during high school. Note that these findings hold true for both the subsets A and B of students. However, there is no significant difference in social activities between the first-gens and non-first-gens during high school. This helps us better understand the real differences in at-risk factors between first-gens and non-first-gens.
5.5 Analysis

Table 5.4: Comparison of risk factors between first-generation and non-first-generation students using T-test. The scores of risk factors are normally distributed (p > 0.05 using KS-test). Full statistical results on subset A and subset B can be found in the supplementary document.

<table>
<thead>
<tr>
<th>risk factors</th>
<th>first-gen mean(std)</th>
<th>non-first-gen mean(std)</th>
<th>t-test p-value</th>
<th>significance holds true on subsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>SES</td>
<td>1.70(0.72)</td>
<td>3.72(0.91)</td>
<td>&lt; 0.001***</td>
<td>⬡ ⬡</td>
</tr>
<tr>
<td>Lifestyle</td>
<td>2.87(0.64)</td>
<td>3.55(0.74)</td>
<td>&lt; 0.001***</td>
<td>⬡ ⬡</td>
</tr>
<tr>
<td>Sociability</td>
<td>3.53(0.87)</td>
<td>4.13(0.64)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* p-value <= 0.05; ** p <= 0.01; *** p <= 0.001, bold if FDR-adjusted p <= 0.05

Comparing Mental Health of First-gens and Non-first-gens.

Next, we investigate whether the challenges first-generation students face as they adapt to college life influences their mental health. Specifically, we compare mental health (using PHQ-4) across each term. PHQ-4 is used clinically to assess depression and anxiety. We use a contrast variable to indicate first-generation students and non-first-generation students. Group differences are examined using linear regression models [207] with the contrast variable as a predictor. Given the repeated measures design of the PHQ-4 data, we use mixed effects models with repeated PHQ-4 scores nested within an individual, and the group contrast is treated as the fixed effect.

Table 5.5 shows the relationship between PHQ-4 scores and the two student groups (i.e., first-gen or non-first-gen). Note that, for the fall term, we only use EMA data collected during mid-term when all the participants had been fully enrolled in the study. The mean and standard deviation of the PHQ-4 score (averaged participant-wise) for both the first-generation and non-first-generation students are also included to provide additional insight into the data distribution. We report the coefficient (β),
5.5 Analysis

Table 5.5: Comparison of PHQ-4 scores between first-generation and non-first-generation students using mixed effects with repeated PHQ-4 scores nested within the individual, and the group contrast (1: 1st-gens & 0: non-first-gens) treated as a fixed effect. First-generation students had significantly higher depression scores during the fall term only. The coefficients $\beta$ decrease as the year progresses, possibly indicating that first-generation students adapt to overcome initial risk factors. Full statistical results on subset A and subset B can be found in the supplementary document.

<table>
<thead>
<tr>
<th>term</th>
<th>first-gen mean(std)</th>
<th>non-first-gen mean(std)</th>
<th>mixed effects result</th>
<th>significance holds true on subsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>fall</td>
<td>3.29(2.12)</td>
<td>2.26(1.74)</td>
<td>$\beta = 1.12, SE = 0.39, p = 0.004^{**}$</td>
<td>$\beta = 1.12, SE = 0.39, p = 0.004^{**}$</td>
</tr>
<tr>
<td>winter</td>
<td>2.51(2.31)</td>
<td>2.21(1.97)</td>
<td>$\beta = 0.45, SE = 0.36, p = 0.21$</td>
<td>$\beta = 0.45, SE = 0.36, p = 0.21$</td>
</tr>
<tr>
<td>spring</td>
<td>2.38(2.25)</td>
<td>2.30(1.99)</td>
<td>$\beta = 0.30, SE = 0.45, p = 0.50$</td>
<td>$\beta = 0.30, SE = 0.45, p = 0.50$</td>
</tr>
</tbody>
</table>

* p-value <= 0.05; ** p <= 0.01; *** p <= 0.001; bold if FDR-adjusted p <= 0.05

the standard error of the estimated coefficient (SE), and its significance as a result of mixed effects models. First-generation students have a higher PHQ-4 score (i.e., lower mental health status) than non-first-generation students. However, we observe that first-generation students’ higher PHQ-4 scores are statistically significant only in the fall term ($\beta = 1.12, p = 0.004$), and the differences in PHQ-4 scores between the two groups of students during the winter and spring terms is no longer significant. The finding is valid on both subsets A and B. This result supports our hypothesis that the first-generation students adapt to overcome initial risk factors as the year progresses.

Examining the Correlation between Risk Factors and Mental Health.

We examine how pre-college risk factors are associated with mental health. PHQ-4 scores are analyzed using mixed effects models with the repeated PHQ-4 nested within
an individual, and each risk factor is treated as a fixed effect. As shown in Table 5.6, we find that higher SES is not significantly associated with better mental health (i.e., a lower PHQ-4 score), especially as the year progresses (β = −0.15, p = 0.18 for the fall term and it continues to be insignificant throughout the other terms). Better support is also not significantly associated with better mental health. However, better lifestyle (β = −0.71, p < 0.001 for the fall term and continues to be significant beyond the fall term) and sociability (β = −0.69, p < 0.001 in the winter term and β = −0.62, p = 0.001 in the spring term) indicates better mental health during the first year at college. It is important to note that many of these findings hold true for both subsets we perform tests on.

Table 5.6: Associations between risk factors and mental health using mixed effects models with the repeated PHQ-4 nested within the individual, and each risk factor (Likert scale from 1 to 5) treated as a fixed effect. Better lifestyle and sociability are associated with better mental health during the first year at college (i.e., lower PHQ-4 score). Full statistical results on subset A and subset B are available in the supplementary document.

<table>
<thead>
<tr>
<th>risk factors</th>
<th>term</th>
<th>association with PHQ-4 based on mixed effects models</th>
<th>significance hold true on subsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>SES</td>
<td>fall</td>
<td>β = −0.15, SE = 0.12, p = 0.18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>winter</td>
<td>β = −0.01, SE = 0.11, p = 0.89</td>
<td></td>
</tr>
<tr>
<td></td>
<td>spring</td>
<td>β = 0.00, SE = 0.13, p = 0.99</td>
<td></td>
</tr>
<tr>
<td>Lifestyle</td>
<td>fall</td>
<td>β = −0.71, SE = 0.16, p &lt; 0.001***</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>winter</td>
<td>β = −0.84, SE = 0.14, p &lt; 0.001***</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>spring</td>
<td>β = −0.89, SE = 0.17, p &lt; 0.001***</td>
<td>A</td>
</tr>
<tr>
<td>Sociability</td>
<td>fall</td>
<td>β = −0.51, SE = 0.17, p = 0.002**</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>winter</td>
<td>β = −0.69, SE = 0.15, p &lt; 0.001***</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>spring</td>
<td>β = −0.62, SE = 0.18, p = 0.001***</td>
<td>A</td>
</tr>
<tr>
<td>Support</td>
<td>fall</td>
<td>β = −0.20, SE = 0.14, p = 0.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>winter</td>
<td>β = −0.20, SE = 0.13, p = 0.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>spring</td>
<td>β = −0.23, SE = 0.16, p = 0.15</td>
<td></td>
</tr>
</tbody>
</table>

* p-value <= 0.05; ** p <= 0.01; *** p <= 0.001, bold if FDR-adjusted p <= 0.05
5.5 Analysis

5.5.2 Q2: What are the Key Behavioral Differences between First-generation and Non-first-generation Students across Each Term as They Progress through Their First Year?

In this section, we examine the differences in behavioral sensing data collected by smartphones. Behavioral sensing data and inferences are computed daily. We analyze the data using mixed effects models nested within an individual and treat the first-gens/non-first-gens group contrast as the fixed effect. For the fall term, statistical analysis is based on data collected after midterm, when participants are completely enrolled. In addition, features with strongly right-skewed distributions (e.g., time spent at Greek houses and time spent at the gym) are log-transformed, resulting in approximately normally distributed data.
Table 5.7: Comparison of behavioral features between first-generation and non-first-generation students. Behavioral sensing data and inferences are computed daily. They are analyzed using mixed effects models nested within individual and the first-gens/non-first-gens group contrast (1: 1st-gens & -1: non-first-gens) treated as the fixed effect. Associations between behavioral features and risk factors are examined using mixed effects models with the daily behavioral features nested within the individual. Each risk factor (Likert scale from 1 to 5) is treated as a fixed effect. The statistical analysis for the fall term is based on data collected after midterm, when participants were completely enrolled. Full statistical results on two subsets are available in the supplementary document.

<table>
<thead>
<tr>
<th>behavior</th>
<th>first-gen mean(std)</th>
<th>non-first-gen mean(std)</th>
<th>mixed effects result</th>
<th>significance hold true on subsets</th>
<th>association with risk factors (p &lt;= 0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Duration at studying places (hr)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fall</td>
<td>7.08(5.20)</td>
<td>3.71(3.66)</td>
<td>$\beta = 1.57, SE = 0.51, p = 0.002^*$</td>
<td>false</td>
<td></td>
</tr>
<tr>
<td>winter</td>
<td>6.28(5.22)</td>
<td>3.53(3.79)</td>
<td>$\beta = 1.40, SE = 0.47, p = 0.003^*$</td>
<td>false</td>
<td></td>
</tr>
<tr>
<td>spring</td>
<td>5.70(4.50)</td>
<td>3.02(3.41)</td>
<td>$\beta = 1.13, SE = 0.41, p = 0.006^*$</td>
<td>false</td>
<td></td>
</tr>
<tr>
<td><strong>Class attendance rate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fall</td>
<td>0.69(0.15)</td>
<td>0.69(0.17)</td>
<td>$\beta = 0.00, SE = 0.08, p = 0.978$</td>
<td>false</td>
<td></td>
</tr>
<tr>
<td>winter</td>
<td>0.62(0.17)</td>
<td>0.59(0.18)</td>
<td>$\beta = 0.02, SE = 0.00, p &lt; 0.001^{***}$</td>
<td>true</td>
<td></td>
</tr>
<tr>
<td>spring</td>
<td>0.62(0.18)</td>
<td>0.64(0.18)</td>
<td>$\beta = -0.02, SE = 0.00, p &lt; 0.001^{***}$</td>
<td>true</td>
<td></td>
</tr>
<tr>
<td><strong>Duration at food place (hr)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fall</td>
<td>0.91(0.69)</td>
<td>1.21(0.64)</td>
<td>$\beta = -0.32, SE = 0.14, p = 0.02^*$</td>
<td>false</td>
<td></td>
</tr>
<tr>
<td>winter</td>
<td>0.92(0.62)</td>
<td>1.05(0.65)</td>
<td>$\beta = -0.26, SE = 0.13, p &lt; 0.06$</td>
<td>true</td>
<td></td>
</tr>
<tr>
<td>spring</td>
<td>0.84(0.59)</td>
<td>1.07(0.57)</td>
<td>$\beta = -0.13, SE = 0.12, p &lt; 0.29$</td>
<td>true</td>
<td></td>
</tr>
<tr>
<td><strong>Sleep duration regularity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### 5.5 Analysis

<table>
<thead>
<tr>
<th>Season</th>
<th>Variable</th>
<th>Fall</th>
<th>Winter</th>
<th>Spring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Effect Size</strong></td>
<td></td>
<td><strong>Effect Size</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Beta</strong></td>
<td><strong>SE</strong></td>
<td><strong>Beta</strong></td>
<td><strong>SE</strong></td>
</tr>
<tr>
<td><strong>Physical activity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fall</td>
<td>-3.50(2.26)</td>
<td>-2.24(1.60)</td>
<td>$\beta = -1.00, SE = 0.06, p &lt; 0.001^{***}$</td>
<td></td>
</tr>
<tr>
<td>winter</td>
<td>-2.33(1.97)</td>
<td>-2.35(1.16)</td>
<td>$\beta = -0.58, SE = 0.03, p &lt; 0.001^{***}$</td>
<td></td>
</tr>
<tr>
<td>spring</td>
<td>-3.01(1.91)</td>
<td>-2.34(1.32)</td>
<td>$\beta = -0.63, SE = 0.03, p &lt; 0.001^{***}$</td>
<td></td>
</tr>
<tr>
<td><strong>Duration at Greek house (log-transformed)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fall</td>
<td>-20.14(3.93)</td>
<td>-17.97(3.42)</td>
<td>$\beta = -1.07, SE = 0.72, p = 0.13$</td>
<td></td>
</tr>
<tr>
<td>winter</td>
<td>-19.63(2.47)</td>
<td>-17.58(2.74)</td>
<td>$\beta = -1.56, SE = 0.57, p = 0.006^{**}$</td>
<td></td>
</tr>
<tr>
<td>spring</td>
<td>-19.80(2.10)</td>
<td>-17.45(3.26)</td>
<td>$\beta = -2.01, SE = 0.68, p = 0.003^{**}$</td>
<td></td>
</tr>
<tr>
<td><strong>Duration at gyms (log-transformed)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fall</td>
<td>-20.46(3.74)</td>
<td>-17.60(5.02)</td>
<td>$\beta = -2.65, SE = 1.05, p = 0.01^{**}$</td>
<td></td>
</tr>
<tr>
<td>winter</td>
<td>-19.25(1.98)</td>
<td>-15.96(5.35)</td>
<td>$\beta = -4.20, SE = 1.05, p &lt; 0.001^{***}$</td>
<td></td>
</tr>
<tr>
<td>spring</td>
<td>-20.18(2.62)</td>
<td>-17.07(4.18)</td>
<td>$\beta = -2.69, SE = 0.89, p = 0.002^{**}$</td>
<td></td>
</tr>
</tbody>
</table>

* p <= 0.05; ** p <= 0.01; *** p <= 0.001; bold if FDR-adjusted p <= 0.05.

SES: SES; lifestyle: lifestyle; sociability: sociability; support: support; ↑: positive association; ↓: negative association;
5.5 Analysis

Table 5.7 compares the behavior of first-generation and non-first-generation students based on mobile sensing data. We observe that first-generation students spend more time in the study areas than non-first-generation students. This is consistent with the finding of [322], which indicates that first-generation students are more likely to use libraries and study spaces. Class attendance is an important metric when assessing college life. First-generation students appear to have a higher rate of winter class attendance than non-first-generation students, but a lower rate during spring classes. Prior studies suggest that such differences could be because of perceived classroom competitiveness, which contributes to differences in engagement, attendance and retention in courses [52]. While first-generation students seem to spend less time in the food area than non-first-generations in the fall, the difference diminishes and becomes insignificant during the winter and spring terms. First-generation students appear to have less consistent sleep duration than non-first-generation students. It could be that they are still adjusting to new academic demands, living arrangements, and social demands and thus end up cutting back on one of the few things they can control – the amount of sleep they get [72], resulting in poorer sleeping habits. First-generation students also spend less time partying at Greek houses. This may bolster the argument that first-generation students are less adept at cultivating social capital – privileged knowledge, resources, and information acquired through social networks – than non-first-generation students [39]. In addition, first-generation students appear to be slightly less physically active than non-first-generation students, although this difference is not significant in the fall and winter; they also spend less time at gyms. This is consistent with the finding in [247], which indicates that first-generation students are generally less physically active. Recall our finding from earlier
5.5 Analysis

that first-generation students are more at risk for lifestyle-related risk factors. Thus, first-generation students being less physically active ties back to such risk-proneness [97, 96]. Another possibility is that they are spending less time exercising due to other demands, such as working at the college, course load, surroundings and schedule [229]. Many first-generation students at Dartmouth work during the term on campus, doing various jobs to earn additional money to support living expenses.

Table 5.7 also illustrates the relationship between such distinguishing behaviors and risk factors, based on mixed effects models with the daily behavioral features nested within an individual and each risk factor (Likert scale from 1 to 5) treated as a fixed effect. We find that socioeconomic status (SES) is inversely associated with time spent at study locations on campus. Class attendance rate is positively related to a better lifestyle, sociability and support. The duration of time spent at food halls is positively associated with SES. Regularity in sleep duration is linked with improved scores across all four risk factors. Physical activity is positively related to SES. The duration of time spent at gyms relates to a healthier lifestyle. On the other hand, the duration spent at Greek houses is unrelated to any of the risk factors.

5.5.3 Q3: What Behaviors Do First-generation Students Exhibit That Are Associated with Better Mental Health? And Are These Behaviors Similar to Those Exhibited by the Non-first-generation Students?

To be able to answer this question, we build a LASSO [324] regression model to predict the mental health of both student groups using the sensing data. The LASSO
5.5 Analysis

(Least Absolute Shrinkage and Selection Operator) regression performs feature selection during the fitting process, excluding less important features concerning PHQ-4. We use the average score of PHQ-4 across a term for each participant as ground truth, indicating whether the subject is depressed or anxious during the term. Behavioral features are treated similarly, with all features aggregated into a term-long daily average for each participant. The behavioral features are then standardized across all subjects. Next, we analyze the coefficients of two linear models to see how each group’s behavior relates to outcomes. For reliability, in Q3 and Q4, we skip the fall term from this analysis, where participants have different lengths of sensor data due to the rolling enrollment process. Table 5.8 shows coefficients of LASSO regression for estimating PHQ-4 of first-generation and non-first-generation students. Each variable is scaled to aid convergence and obtain regression coefficients that could be compared for relative importance.

For non-first-generation students, we observe very similar linear models for both the winter and spring terms, as the LASSO model selects common features with similar coefficients for both terms, as shown in Table 5.8. In the case of non-first-generation students, the regression model mostly relies on time spent at Greek houses, time spent at other students’ dorms, physical activity, class attendance, regularity of sleep onset and circadian rhythm of physical activity. The time spent partying at Greek houses appears to be negatively associated with PHQ-4, indicating better mental health. Surprisingly, higher class attendance is also associated with better mental health. We also observe that regularity of sleep, physical activity, eating, and location routines appear negatively associated with PHQ-4. This indicates that the regularity of routines in life relates to better mental health.
Table 5.8: The coefficients of LASSO (Least Absolute Shrinkage and Selection Operator) regression for estimating PHQ-4.

<table>
<thead>
<tr>
<th></th>
<th>non-first-generation</th>
<th>first-generation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>winter</td>
<td>spring</td>
</tr>
<tr>
<td>Time spent at Greek houses</td>
<td>-0.35</td>
<td>-0.57</td>
</tr>
<tr>
<td>Time spent at own dorm</td>
<td>-0.19</td>
<td></td>
</tr>
<tr>
<td>Time spent at other’s dorm</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>Time spent at study places</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time spent at gyms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of location visited</td>
<td>0.51</td>
<td>0.68</td>
</tr>
<tr>
<td>Phone unlock duration at study places</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sleep offset time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sleep onset time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class attendance</td>
<td>-0.20</td>
<td>-0.42</td>
</tr>
<tr>
<td>Regularity of sleep onset</td>
<td>-0.14</td>
<td>-0.04</td>
</tr>
<tr>
<td>Regularity of sleep duration</td>
<td>-0.29</td>
<td>0.09</td>
</tr>
<tr>
<td>Circadian rhythm of physical activity</td>
<td>-0.44</td>
<td>-0.65</td>
</tr>
<tr>
<td>Eating regularity</td>
<td>-0.06</td>
<td>0.68</td>
</tr>
<tr>
<td>Location regularity</td>
<td>-0.34</td>
<td>-0.91</td>
</tr>
</tbody>
</table>

Physically active duration (e.g., walking, running) has a positive standardized coefficient. This is an interesting observation that might lead us to falsely believe that more physical activity is associated with worse mental health, which goes against the findings of prior research [342, 346]. However, such coefficients can also be influenced by the correlation between features. When we check the correlation between variables, we find that for non-first-generation students, time spent at Greek houses is positively related to the physically active duration ($r=0.38$, $p<0.001$ in the winter term and $r=0.4$, $p<0.001$ in the spring term). Therefore, it appears that time spent at Greek houses may have absorbed some of the variances associated with the physically active duration. A deeper dive into the results (as shown in Table 5.8) shows that when
5.5 Analysis

LASSO does not select the time spent at Greek houses as a predictor, physically active
duration has a negative standardized coefficient, thus indicating the more physically
active first-generation students are the better their mental health.

Among first-generation students, we observe very distinct linear models across
the winter and spring terms. In fact, there are only two features with a common
direction shared by models during these terms; that is, the phone unlock duration at
study places and location regularity between the same days of weeks. The phone un-
lock duration at study places has a positive coefficient, indicating that the increased
phone usage at study areas relates to poorer mental health. This finding is in line
with prior findings [346] where the authors hypothesize that phone usage in the class-
room and study places is a potential indicator of a student’s diminished ability to
concentrate – one of the depressive disorder symptoms. Our analysis suggests that
such a connection seems more obvious among first-generation students. Similar to the
non-first-generation students, we observe location regularity is negatively associated
with PHQ-4 for first-generation students. This might mean that a general regularity
in semantic location routines between days possibly relates to better mental health.
However, we observe some distinct patterns associated with first-generation students;
for example, during the winter term, regularity of visits to food places has a positive
standardized coefficient. Gathering at a common dining place may not just be related
to the consumption of meals but other hidden activities such as sitting and chatting
with friends [310], or in the case of many of our first-generation students working
at cafeterias on campus. It may be possible that eating regularity may have a dif-
ferent contextual meaning for the first-generation students and non-first-generation
students, especially at the early stages of their college life. We also observe that
5.5 Analysis

the time spent at Greek houses is selected as a predictor for the spring term only, possibly because the first-generation students adapt and participate in more social events at Greek houses as the year progresses (see Table 5.7). Visiting more locations on campus is associated with better mental health during the spring term but not during the winter term, possibly due to the seasonal effect (i.e., cold weather). Overall, we notice only a few commonalities in the linear models of first-generation and non-first-generation students. First-generation students have a range of unique behavioral predictors associated with their mental health. Such predictors appear to even change over time. As a result, first-gens start to overcome initial challenges, and by navigating various academic and social demands, they begin to adapt to college life positively.

5.5.4 Q4: How Can We Predict the Mental Well-being of At-risk Students More Accurately, with a Particular Focus on the First-generation Students?

Inferring mental health from passive sensing is not a novel idea, and machine learning has been widely used to accomplish this [329, 346]. Prior research has employed both classical machine learning [329, 346] and deep learning techniques [6, 300, 373]. However, previous research focuses exclusively on model performance over the entire participant population, without particularly accounting for a distinct subpopulation. Our exploratory research suggests that generic models found in the existing literature that are trained without taking into account the variations between first-generation and non-first-generation students are not optimal in assessing the mental health of first-generation students. However, training models individually between the two
5.5 Analysis

groups is not desirable because we lack adequate first-generation samples and because
this strategy precludes the two groups from learning common information. As a result,
we require an architecture that can be used to train the model on the complete dataset
while also taking into account the distinctions between first-generation and non-first-
generation students, as previously stated.

Data Preprocessing

The sensing data of one participant in a term can be viewed as a multivariate time
series with a length equal to the number of days in a term. For each day in the term, we
include the 83 features outlined in Section 5.4.3. Next, we standardize these features
within and between participants. We employ robust scaling to deal with outliers,
where the values of each variable are subtracted from their median and divided by
the interquartile range (IQR) – the difference between the 75th and 25th percentiles.
After handling outliers, we standardize features in two ways. First, we perform within-
person standardization using all days in a term to capture intra-person variability
in behaviors. Second, we standardize features across participants to capture inter-
person or population-level information to improve model diversity. Ultimately, the
two standardized feature sets are concatenated to form a time sequence with 166
features for each day. Each term is considered as a separate data point, with the
dimensions (t, d), where t = 70 represents the number of days in a term and d =
166 represents the number of features. As discussed earlier, PHQ-4 is a two-week
screening scale for anxiety and depression. We are interested in a student’s mental
health during the term. When a student has multiple EMAs within a term, researchers
use the average of the responses during the term to indicate how the subject feels
5.5 Analysis

in general throughout the time period [163]. Thus, we use the average value of a student’s PHQ-4 EMAs as the output ground truth. We use a cutoff score of 3 because clinically, PHQ-4 scores greater than 3 are seen in persons with anxiety and depression [189]. Using a cutoff score of 3, 34% of the data samples are positive. To ensure a rigorous evaluation, we set aside data from 20% of students (leave-subjects-out) for testing purposes only (we refer to this as D2) in the following sections. The remaining 80% of the data is used for training and validation (we refer to this as D1).

Baseline: Using Traditional Machine Learning Approaches

We first start with the basics and discuss whether traditional machine learning methods can predict students’ mental health, considering first-generation students. Traditional machine learning approaches are incapable of directly handling time series. As a result, we take an average over the time axis and reduce the dimension of each data point to a vector of length d (where d equals 166, the number of features). Such a method is commonly used in current studies [329, 346]. In addition, two categorical features are added to the input: (1) whether a participant is first-generation, and (2) which term the data point belongs to. Next, we perform a grid search to optimize the model hyper-parameters within a stratified 4-fold leave-n-subject-out cross-validation using D1. We then train a model on the combined training and validation data using the selected super-parameters and evaluate this model on D2. Table 5.9 shows the predictive performance of several traditional machine learning classifiers. While these models can achieve precision scores slightly above 0.7 and an F1 score close to 0.7 across all participants, the predictive performance of these machine learning models for first-generation students is poor, as indicated by lower F1 scores. In addition, the
5.5 Analysis

recall rate is low. This could be because of the nature of such approaches. Traditional machine learning models cannot capture valuable temporal information after averaging over time. In addition, even if we indicate in the input that a student is a first-generation student, such algorithms will not fully exploit this information. For example, linear approaches may include an additional intercept on categorical variables (e.g., being first-generation). In contrast, decision tree-based methods may ignore such categorical features at higher levels if they cannot provide sufficient information gain. As a result, the model is more likely to be trained to fit the majority of students who are not first-generation.

Table 5.9: Predictive performance on PHQ-4 using traditional machine learning approaches. Traditional machine learning approaches fail to make predictions among first generations. Metrics are weighted to account for label imbalance.

<table>
<thead>
<tr>
<th>Machine Learning method</th>
<th>precision (among all participants)</th>
<th>recall (among all participants)</th>
<th>F1-score (among all participants)</th>
<th>F1-score (among 1st-gen)</th>
</tr>
</thead>
<tbody>
<tr>
<td>logistic regression</td>
<td>0.73</td>
<td>0.56</td>
<td>0.67</td>
<td>0.49</td>
</tr>
<tr>
<td>random forest</td>
<td>0.71</td>
<td>0.58</td>
<td>0.65</td>
<td>0.48</td>
</tr>
<tr>
<td>gradient boosting</td>
<td>0.73</td>
<td>0.56</td>
<td>0.69</td>
<td>0.42</td>
</tr>
</tbody>
</table>

**Deep Learning Architecture**

The aforementioned limitations of traditional machine learning models hinder the learning of good time series representations. In contrast, deep learning approaches can learn more sophisticated representations capturing the temporal nature of sensor data. Specifically, we propose a unique multi-task architecture that can leverage auxiliary information about the student’s group (first-gen or not) and term (academic progress/ seasonal information).
According to Section 5.5.4, the dimension of each data input is \((t=70, d=166)\), which represents a student’s overall behavioral pattern over the duration of a term. We first use bidirectional Long Short-Term Memory (LSTM) network \([153]\) to generate the representation of the input data. Bidirectional LSTM connects two hidden layers of opposite directions to improve model performance on sequence classification problems \([297]\). For each data point, we keep all hidden states at each step of the recurrent neural network so that the bidirectional LSTM layer generates an output \(H\) with the shape of \((t = 70, 2 \times \text{hidden units})\). We further add self-attention \([205]\), a technique widely adopted in natural language processing for sentence embedding, to the bidirectional LSTM models to assign different weights to different days in a term emphasizing their importance in predicting depression.

Figure 5.3: Adding two auxiliary output in a neural network architecture where the losses are provided by the labels of whether the student is a first-generation student and the academic/college life progress.
5.5 Analysis

Table 5.10: The details of DNN-1st-gen, DNN-term and DNN-PHQ in the architecture.

<table>
<thead>
<tr>
<th>layer</th>
<th>hidden units</th>
<th>activation function</th>
<th>dropout rate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DNN-1st-gen</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>output from self-attentive bi-LSTM</td>
<td>512</td>
<td>ReLU</td>
<td>0.5</td>
</tr>
<tr>
<td>hidden fully connected</td>
<td>512</td>
<td>ReLU</td>
<td>0.5</td>
</tr>
<tr>
<td>hidden fully connected</td>
<td>128</td>
<td>ReLU</td>
<td></td>
</tr>
<tr>
<td>auxiliary output</td>
<td>1</td>
<td>sigmoid</td>
<td></td>
</tr>
<tr>
<td><strong>DNN-term</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>output from self-attentive bi-LSTM</td>
<td>512</td>
<td>ReLU</td>
<td>0.5</td>
</tr>
<tr>
<td>hidden fully connected</td>
<td>512</td>
<td>ReLU</td>
<td>0.5</td>
</tr>
<tr>
<td>hidden fully connected</td>
<td>128</td>
<td>ReLU</td>
<td></td>
</tr>
<tr>
<td>auxiliary output</td>
<td>1</td>
<td>sigmoid</td>
<td></td>
</tr>
<tr>
<td><strong>DNN-PHQ</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>output from concatenate layer</td>
<td>768</td>
<td>ReLU</td>
<td>0.5</td>
</tr>
<tr>
<td>hidden fully connected</td>
<td>1024</td>
<td>ReLU</td>
<td>0.5</td>
</tr>
<tr>
<td>hidden fully connected</td>
<td>512</td>
<td>ReLU</td>
<td>0.5</td>
</tr>
<tr>
<td>hidden fully connected</td>
<td>256</td>
<td>ReLU</td>
<td>0.5</td>
</tr>
<tr>
<td>hidden fully connected</td>
<td>128</td>
<td>ReLU</td>
<td></td>
</tr>
<tr>
<td>main output</td>
<td>1</td>
<td>sigmoid</td>
<td></td>
</tr>
</tbody>
</table>

To enable the model to pay attention to differences between the first-generations and non-first-generations, we add two auxiliary outputs in our neural network architecture (Figure 5.3). We have observed that (1) students’ behaviors change in different terms, possibly due to mixed factors, e.g., academic demands and seasonal effects (Section 5.3.2), (2) first-generation students have different behaviors compared to the non-first-generations (Section 5.5.2), and (3) the association between the behavior and mental health may be different as the year progresses (Section 5.5.3). Therefore, we use two auxiliary outputs, where the losses are provided by the labels of whether the student is a first-generation student and their academic progress (i.e., which term the student is in). Using these auxiliary outputs, we attempt to con-
5.5 Analysis

textualize mental well-being. We create two independent, fully connected sub-neural networks (DNN-1st-gen and DNN-term as shown in Figure 5.3) to take the output of the self-attentive bidirectional LSTM and generate auxiliary outputs. The details of the DNN-1st-gen and DNN-term, in terms of the hidden layers and number of hidden units, activation functions, dropout rate, and sizes of input and output, are shown in Table 5.10. The output of the self-attentive bidirectional LSTM is the input of DNN-1st-gen and DNN-term. In both DNN-1st-gen and DNN-term, there are two fully connected hidden layers with Rectified Linear Unit activation functions (ReLU). The hidden units of the last hidden layer (before the auxiliary output) of DNN-1st-gen and DNN-term networks are concatenated with the original output from self-attentive bidirectional LSTM and fed into a DNN-PHQ network (see Table 5.10 for details), which generates the main the depression output (i.e., PHQ-4). In this way, the network is forced to learn from the behavioral sensor data with the emphasis on the specific term as well as the student group (i.e., whether the data under consideration is associated with a first-generation student).

Next, we optimize the weights assigned to the main loss and auxiliary losses to improve the network's predictive performance. The loss function for each of the three outputs is binary cross-entropy. Note that the fall term is skipped due to incomplete data from the rolling enrollment process, and thus the term has only two labels. Starting from (1, 0, 0), which means the network is only trained from the main loss (i.e., PHQ-4) and does not back-propagate the loss of the auxiliary output, we gradually reduce the weight of the main output and increase the weights of the auxiliary ones. As we do not have prior knowledge of which of the two auxiliary outputs is more important, we test three different settings of the weights while keeping
the main loss fixed (as noted in the same color in Table 5.11): (1) weight(DNN-1st-gen) > weight(DNN-term), (2) weight(DNN-1st-gen) = weight(DNN-term), and (3) weight(DNN-1st-gen) < weight(DNN-term). To prevent overfitting during training, we stratified D1 data (leave-subjects) into a training (60%) and validation (20%) set. We monitor the loss on the validation set and early-stop the training process if the loss does not decrease after more than ten iterations. Following this, the trained model is evaluated on D2 data. Table 5.11 shows the performance of models with different settings of weights. The results show that the F1-score firstly drops when very little auxiliary loss is added to the network. This might be because the additional auxiliary loss throws off the learning process, as it can not provide enough information because of the small weight. As we increase the weight of the auxiliary loss, the performance increases and reaches the best performance for the following settings: the weight(DNN-PHQ) = 0.7, weight(DNN-1st-gen) = 0.15, weight(DNN-term) = 0.15. The results show that with a similar weight(DNN-PHQ), the network performs best when weight(DNN-1st-gen) is approximately equal to weight(DNN-term), as shown by the dark orange, dark green and dark yellow rows in the table. Finally, the performance drops if the weight of our main output, i.e., when the weight(DNN-PHQ) is too small.

In summary, by having the learning model pay attention to whether students are first-generation or not and the academic term, the neural network improves the F1-score by 0.07, increasing performance from 0.63 to 0.70. Importantly, the proposed architecture overcomes the drawbacks of generic baseline models, which perform poorly for first-generation students. In particular, our model improves the F1-score among the first-generation students by 22%, increasing it from 0.58 to 0.71.
5.5 Analysis

Table 5.11: Predictive performance of depression (PHQ-4) when different weights are assigned to the main loss and the auxiliary losses. Metrics are weighted to account for label imbalance.

<table>
<thead>
<tr>
<th>Deep Learning weight (PHQ, 1st-gen, term)</th>
<th>precision</th>
<th>recall</th>
<th>F1-score</th>
<th>F1-score (among 1st-gen)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 0, 0) – baseline</td>
<td>0.62</td>
<td>0.63</td>
<td>0.63</td>
<td>0.58</td>
</tr>
<tr>
<td>(0.95, 0.05, 0)</td>
<td>0.61</td>
<td>0.63</td>
<td>0.61</td>
<td>0.42</td>
</tr>
<tr>
<td>(0.95, 0.05, 0.05)</td>
<td>0.57</td>
<td>0.59</td>
<td>0.58</td>
<td>0.42</td>
</tr>
<tr>
<td>(0.8, 0.15, 0.05)</td>
<td>0.62</td>
<td>0.63</td>
<td>0.62</td>
<td>0.8</td>
</tr>
<tr>
<td>(0.8, 0.1, 0.1)</td>
<td>0.65</td>
<td>0.66</td>
<td>0.65</td>
<td>0.63</td>
</tr>
<tr>
<td>(0.8, 0.05, 0.15)</td>
<td>0.61</td>
<td>0.64</td>
<td>0.6</td>
<td>0.57</td>
</tr>
<tr>
<td>(0.7, 0.2, 0.1)</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.71</td>
</tr>
<tr>
<td>(0.7, 0.15, 0.15)</td>
<td>0.69</td>
<td>0.7</td>
<td>0.7</td>
<td>0.71</td>
</tr>
<tr>
<td>(0.7, 0.1, 0.2)</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.52</td>
</tr>
<tr>
<td>(0.6, 0.25, 0.15)</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.52</td>
</tr>
<tr>
<td>(0.6, 0.2, 0.2)</td>
<td>0.64</td>
<td>0.66</td>
<td>0.64</td>
<td>0.58</td>
</tr>
<tr>
<td>(0.6, 0.15, 0.25)</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Our approach outperforms existing models, which are biased toward the majority – non-first-generation students – while underperforming the minority first-generation students. Our approach mitigates bias and improves performance simultaneously.

Interpretation

Following modeling, the natural next step is to explain better which features are essential when predicting mental health. We use the permutation feature importance algorithm [109], a model-agnostic interpretation approach that abstracts the explanations from the deep learning model [280] in which the importance of a feature is measured by the increase in the model’s prediction error after permuting the feature. The permutation is repeated 10 times for each feature. Figure 5.4 shows the top 10 most important features. The x-axis displays the importance score, measured
5.5 Analysis

(a) Feature importance in predicting depression (PHQ-4) of non-first-generation students

(b) Feature importance in predicting depression (PHQ-4) of first-generation students

Figure 5.4: Top 10 most important features from deep learning model discovered using permutation feature importance algorithm. The x-axis displays the decrease in the F1-score when a feature is permuted. The y-axis displays the name and index of each of the 166 features.
5.6 Discussion

by the decrease of the F1 score if a feature is permuted. The y-axis shows the feature name and index among the 166 features, with “wi” indicating a within-person standardized feature and “bt” indicating a between-person standardized feature (see Section 5.5.4 for standardizing methods). The error bars show the standard deviation of the scores across the 10 permutations for each feature. We examine the best model as discussed in Table 5.10 and 5.11. We find that time spent at various Greek houses on campus and the time spent at their own dorm are the two most important determinants in predicting mental health among non-first-generation and first-generation students, respectively. Additionally, both student groups share the common top-10 characteristics, such as activities on foot (walking/running), distance traveled around midnight, and phone use at night. Certain features are important for non-first-generation students, such as sleep onset time and sleep duration, which are both standardized among individuals. Sleep features tend to be less critical for first-generation students in comparison to non-first-generation students. The number of distinct locations visited both during the day and at night is also important. Finally, results from model-agnostic interpretation indicate the importance of features in predicting depression but not the direction; that is if the interpretation is a positive or negative association – for example, time spent at Greek Houses or time spent at dorms.

5.6 Discussion

In this section, we discuss the main contribution of our work, the benefit of our deep learning approach, the implication of our results on better supporting first-generation students on college campuses, and finally, the limitations of the First-Gen study.
5.6 Discussion

5.6.1 Contribution of This Work

The First-Gen study aims to paint a digital picture from phone sensor data of the first year at Dartmouth College for a small cohort of first-generation students as they rise from high school. We use mobile sensing to capture objective behavioral data over the entire first year. This research makes several contributions, including systems design of a low burden mobile sensing platform capable of high compliance and operating over an entire year; offering new research insights into risk factors and the mental health of first-generation students; and the development of a novel deep learning architecture to accurately predict mental health with a focus on first-generation students.

From the perspective of systems development, we improve the design of the StudentLife app by creatively integrating VoIP push notifications into the iOS sensing app, ensuring high compliance and low battery cost. Our First-Gen dashboard monitors all students’ data compliance, allowing the research assistants “data sitting” the study to reach out to students instantly when they observe missing data. The two pilot studies to collect user input are important to ensure the robustness of the data collection phase of the study. As a result, we observe significantly higher retention and data quality than in previous long-term mobile sensing research. We believe that other researchers in this field can benefit from our experience and insights. From the perspective of student well-being, we investigate the association between various pre-college risk factors, on-campus behavior and mental health across different terms and breaks during the first year. Although previous studies explored many risk factors associated with first-generation students, they are purely based on surveys and periodic self-reports. We collect 24/7 data across the year, offering unprecedented data and insights in a passive, continuous manner. Prior first-gen studies
5.6 Discussion

do not include on-campus longitudinal sensing, leading them to potentially overemphasize pre-college risk factors while underestimating the adjustments students make as college life progresses. For example, we find that from the second term onward, there are no significant differences in the depression (PHQ-4) scores between first-generation and non-first-generation students. Furthermore, among the factors that first-generation students are at more risk for, according to literature, we found that only lifestyle and sociability continue affecting mental health. Our analysis suggests that rather than simply considering the initial risk factors to estimate mental health, we also need to closely note changes in lifestyle and sociability detected via mobile sensing. Our work goes beyond simply presenting depression prediction performance results and attempts to provide deeper insights through interpretation.

In terms of methodology, we propose a novel deep learning architecture to better predict the mental health of students, particularly first-generation students. Traditional machine learning approaches require us to combine group knowledge with mobile sensing features. These traditional methods do well on average, but fail to offer good performance for the minority group. The deep model presented in this chapter includes group and term information as auxiliary outputs, forcing the model to pay attention to the sensing data of first-generation students. We show that by adjusting the weights associated with primary loss and auxiliary losses, we can boost the predictive performance of the group of interest (i.e., first-gens) while maintaining the performance of the broader population. We hypothesize that researchers can replace the auxiliary outputs with other factors of interest, allowing this methodology to be potentially extended to other problems.
5.6 Discussion

5.6.2 The Benefit of Our Deep Learning Approach

There is a debate about why we need deep models when other traditional machine learning models are more interpretable and less of a black box. Although deep learning by itself does not provide interpretable features and insights, we can employ other tools such as model-agnostic interpretation approaches, such as those described in Section 5.5.4, to tease apart essential insights from the machine learning model. We find model-agnostic interpretation beneficial for longitudinal time-series behavioral data from phones. Traditional machine learning algorithms struggle with time-series mobile sensing data, where researchers engineer hand-crafted features or aggregates – statistical features on each time-series data (e.g., average, percentile, maximum, minimum, etc.) – resulting in loss of information and an explosion of highly correlated features. There is no easy way to assign weights to different days in the time series because we do not know which day is more relevant to predicting a specific outcome. Deep learning algorithms, such as LSTM, are adept at handling such time series challenges.

The use of self-attention enables learning models to learn the weights of days from the sensor data directly, avoiding the need for prior information. Figure 5.5 shows an attention vector indicating the weights of different days during a term to predict depression (PHQ-4) for a specific student. In this case, the behavior before and after the midterm period is more important, maybe because other students exhibit similar changes in behavior during midterms, making data collected during the midterm less valuable in distinguishing overall good or bad mental health during the term. Non-deep learning models implicitly treat the whole term evenly. The use of self-attention in this regard improves model generalizability because self-attention can
5.6 Discussion

learn contextual cues from a specific population under study and their experiences. For example, consider the case of different schools having different schedules. A deep model can be trained and customized to the context of a specific school from sensing data without being bound by certain statistical aggregation of the data.

Finally, deep learning allows us to treat the information of interest as auxiliary output rather than input. Such an approach provides flexibility, as the model can be modified to perform well for a separate task as well.

![Figure 5.5: The attention vector indicates the weights of different days during a term in predicting depression (PHQ-4) from sensing data from a person.]](image)

5.6.3 College Support for First-generation Students

Today there is a data and a knowledge gap at colleges supporting at-risk students such as first-gens. Professors have little or no visibility of struggling students outside of academics. Student deans get to know students more personally and may have better insights into how students are coping. Wellness programs inside the administrations set the policy and adopt programs to advance the health and safety of students. Health centers on campuses, both on the clinical and counseling side, are on the
5.6 Discussion

front line when dealing with the ebb and flow of mental health across the school year. Institutional statistics collected each year indicate that the trajectory of mental health across the country is heading in the wrong direction. We observe a disconnect between all the stakeholders in that none has a complete view of the dynamics of students’ mental health. We believe the missing link or the glue involves characterizing mental health using mobile sensing.

Compared to other students, first-generation students arrive at college less prepared and with fewer resources and coping skills. As a result, first-gens struggle with the social, financial and academic demands they confront. For this reason, many colleges have active first-generation programs to help students navigate college. Dartmouth College, for example, has a month-long summer orientation program between high school and college for admitted first-generation students. During orientation, we teach mini-classes and offer skill-building workshops, self-assessment tools, and goal-setting sessions. The goal of the orientation is to Bootcamp their first year, where students learn about the pitfalls, hear about coping skills, talk to faculty and begin to forge a strong peer network that will help carry them through the college years.

Beyond college programs for first-gens, researchers are making progress in designing new mental health technologies to support college students. However, we are in the early stages of this development. A recent study presents a series of design activities conducted with college students and emphasizes the social ecosystems and social support networks in a college student’s life [196]. We believe StudentLife and the First-Gen study could potentially open up new insights and ideas for moving things forward. For example, new forms of intervention may emerge to help keep students healthy and on track academically. Combining behavioral sensing from phones and
wearables with new intervention systems is an open and important area of research.

In our study, we train a model using mobile sensing data and self-reported EMAs. After training the network using the available labels and validating the model’s generalizability, we might eliminate the need for self-reports/EMAs, as the model should be capable of estimating the mental health of new individuals not in the training data. The results discussed in this chapter provide objective insights and, as a next step, could provide possible suggestions to first-gens associated with alleviating poor behavioral patterns by promoting behaviors that lead to better mental health. With such an AI-based sensing and intervention system, risks would be detected early. More importantly, we envision personalized and accurate interventions that provide suggestions based on behavioral sensing data.

Although our work is a proof-of-concept, we believe that the method presented in the chapter can be scaled to a larger population with ease. First, it could target individuals or subgroups (say, first-gens) whose risk of developing depression is higher based on the risk factors. Second, through early detection, we can provide help and guidance to them to ensure that they have a better college experience. And finally, with the understanding of what works and what does not in helping students, we will be able to help reduce the future recurrence of poor behaviors that negatively impact mental health. Future mental health sensing and intervention platforms will offer alerts, guidance, and tips anytime and anywhere, which may be ideal for students who have trouble with in-person appointments. The machine learning model itself can be run offline or on-device to help protect the privacy of students. The modern phones are equipped with neural processing units and high-efficiency edge computing hardware that makes the model training and prediction fast and scalable. We can also
5.6 Discussion

perform the preprocessing either locally or leverage edge or cloud computing. The future of mental health sensing will combine phones and wearables to offer higher performance prediction and personalized sensing and intervention. Such a combined platform would integrate physiological signals, behavioral sensing, and contextual insights (i.e., around the student).

5.6.4 Limitations

Although our study offers many interesting insights into the behavior of first-gen students, it also comes with some limitations that should be noted. First, our study is conducted at a single university. Therefore, our findings may not be consistent with first-generation students in another university. We do not make any claims of generalization of our results. We can say that we found important differences between first-gens and non-first-generation students at Dartmouth College. However, we believe that our methodology has the potential to be generalized, as we discussed in Section 5.6.2. Only through careful replication and reproducibility will we be able to understand best if our results offer general insights.

Second, we enlisted as many students as possible as a university research team, given staffing and funding limitations. However, we must acknowledge that the small size of the first-generation student population may result in false discovery and overfitting. To limit the likelihood of drawing an inaccurate result, we employ rigorous statistical methods (e.g., splitting the data into two random, non-overlapping subsets, FDR and leave-subjects-out test group) to demonstrate the potential of such an approach. Even so, the result and approach must be retested with a bigger sample to determine its generalizability.
5.6 Discussion

Another limitation is the reliability of the iOS/Android activity recognition APIs used to infer physical activities. Since different companies manufacture Android devices, mobile phones may have a vendor-specific hardware sensor and software implementation that might lead to some errors in activity recognition. Additionally, we recognize that the recognition algorithm may evolve across iOS/Android versions. All of these factors would result in inconsistency in activity recognition. Although Apple and Google assert that they ensure accuracy through training on vast amounts of data, it is quite difficult for us to quantify any inherent error in these measurements. This may continue to be a restriction for all studies based on smartphone and wearable sensing, as we have yet to find literature that supports this inaccuracy and resolves any problems associated with it.

Furthermore, even though smartphone sensing can capture various behaviors, it is still limited in capturing very fine-grained physical activities. For instance, students may set their phones down while running on a treadmill, in which case smartphone sensors would be unable to recall this physical activity (on-campus location can serve as a better proxy). In general, one can expect the physical activities (walking / running / cycling / in-vehicle / sedentary) to be more accurate outdoors when participants carry their phones on their bodies and less accurate indoors when participants are more likely to place their phones on a table. In other studies [290, 239, 220], researchers examine alternative sensing modalities (e.g., wearables, in-house sensors) for more precise activity identification to circumvent the limitation of smartphone sensing. However, we did not pursue this course of action in our study because we are focused on better understanding how effectively a low-cost tool, such as StudentLife, predicts the mental health of at-risk students.
5.7 Conclusion

In this chapter, we used the StudentLife mobile sensing platform to track the behavior of N=180 first-year students at Dartmouth college for an entire year. We assessed the risk factors of all students focusing specifically on first-generation students who are more at risk than other groups. We considered various factors associated with first-year students’ risk, including socioeconomic status, lifestyle, sociability and support network. We discussed the First-Gen year-long study to investigate how first-generation students’ behaviors are associated with their depression and anxiety scores. We observed differences in sensed behaviors of students across each term and academic break. We also reported on behavioral differences between first-generation and non-first-generation students. We designed a novel deep neural network architecture that can learn more informative insights from the behaviors of first-generation students, providing more accurate mental health predictions than generic models. We are currently following the first-generation students in this study across their remaining college years and hope to report our findings in the future after they graduate.
Chapter 6

The Need to Adapt: Mitigating Model Degradation in Longitudinal Human Behavioral Sensing Studies

6.1 Introduction

Depression, formally defined as major depressive disorder (MDD), is a prevalent and significant mental health issue that affects individuals all over the world. A recent survey in the United States found that about 21.0 million adults in the country have had at least one major depressive episode. This number represented 8.4% of all U.S. adults. Persons between the ages of 18 and 25 had the highest prevalence of major depressive episodes, at over 17% [254]. More than 66% of college students in the United States reported experiencing excessive anxiety within the past year, and more than 46% reported being so depressed that it was difficult to function[1]. Moreover, 27% of undergraduates were diagnosed or treated for anxiety, depression,
or both, according to the same report. Identifying early warning signs of depression could mitigate or prevent major depression disorder’s negative consequences [50]. In the realm of mobile sensing and intervention, the past decade has witnessed a significant amount of research activity [20]. However, a notable limitation in this field is the prevalence of short-term studies. Conducting longitudinal studies using mobile behavior sensing is time consuming and expensive. Few real-world datasets are available and there has been limited exploration in understanding the degradation of behavioral human models over time, or creating predictive models that maintain their effectiveness and robustness over extended periods. For example, a behavioral model trained on students prior to COVID-19 would likely significantly underperform if not adapted to be responsive to potential impacts and changes brought about by the pandemic (e.g., campuses closing, quarantine, limits on social interaction, remote learning) – or for that matter, behaviors once perturbed by event such as COVID may return to baseline or not (i.e., the new normal).

Over the course of time, it is possible for the distribution of the data being processed by the model to undergo alterations. This can give rise to a scenario in which the underlying assumptions of the model regarding the data are no longer valid, resulting in a deterioration of its performance (so-called data drift). Sometimes, the underlying relationships within the data change over time. This can be attributed to alterations in user concept, societal circumstances, or other variables that were not originally incorporated into the model’s training (so-called concept drift). They are two primary factors that contribute to the occurrence of model degradation, i.e., the decline in the performance of a machine learning model over time.

In this chapter, we collected mobile sensing data and weekly mental health self-
6.2 Related Work

reports from 215 first-year undergraduate students in the United States throughout their four-year university life. We categorized academic terms and holidays into different phases to reflect the changes experienced by individuals in the study, such as the sudden onset of the pandemic and more gradual changes like aging. We investigated the following research questions sequentially:

- (Q1) Is there evidence of domain drift in the context of assessing depression through behavioral sensing over the course of college life?
- (Q2) Does domain drift cause significant model degradation?
- (Q3) Can we use adaptation techniques to counteract this model degradation over time?

Our work aims to enhance the resilience of models to change, examining how the models would accommodate various real-world events that individuals may encounter in their lives.

6.2.1 Mobile Sensing and Mental Health

Mental health sensing is increasingly using mobile devices and wearables for passive, real-time monitoring of individuals in various settings [240]. Presently, advanced sensors in mobile phones and wearable devices facilitate the unobtrusive collection of a wide array of behavioral data without requiring active participation from users. Recently, researchers have started investigating the replication of studies and observed
6.2 Related Work

huge performance gap caused by the distribution shift in different datasets [367]. However, research tracking model performance on the same individuals over significant periods, such as the entirety of a four-year college experience, remains limited. Consequently, researchers have yet to determine the best practices for maintaining a mental health sensing system. This includes understanding the timing of model degradation, the extent to which domain adaptation techniques can mitigate this degradation, and identifying when it becomes necessary to collect additional self-reported mental health labels for model updating.

6.2.2 Domain Adaptation

Domain adaptation is a specific case of transfer learning [185]. Transfer learning is the learning situation in which a model is trained on a source domain or task and then evaluated on a separate but similar target domain or task, where either the tasks or domains (or both) differ [354, 359]. It is a common practice in deep learning to use pre-trained models as a starting point for computer vision and natural language processing tasks, due to the vast computational and time resources required to develop neural network models for these problems and the enormous skill gains that these models provide on related problems [302, 313].

In domain adaptation, the source and target tasks are same, but the domain varies [359]. A domain comprises a feature space and the distribution of features within a dataset. Researchers are especially interested in unsupervised domain adaptation (UDA), in which a model is trained on labeled data from a source domain and then moved to a target domain without any labeled data in the target domain [258, 275, 215]. UDA is beneficial when gathering labels is costly, laborious, or needs specialist
In recent years, researchers have presented a vast array of domain adaptation techniques. Earlier UDA approaches [331, 212] included Maximum Mean Discrepancy (MDD) loss into a neural network’s loss function in order to reduce domain discrepancy. MMD is a nonparametric distance metric between two probability distributions. Recently, UDA has benefited from the application of domain-adversarial training of neural networks [120]. In domain-adversarial learning, a feature extraction network is trained to generate features that are indistinguishable to a domain classifier and highly task discriminative to a task classifier. It is hypothesized that, after training, the features are domain-invariant. UDA has been effectively used in numerous fields, including Natural Language Processing (NLP) [54, 209, 370], object recognition [135, 369], and image comprehension [327, 376, 299], etc.

6.3 Dataset

The absence of longitudinal tracking and discussions on model durability in mental health sensing in previous research may stem from the cost implications of conducting multi-year studies and the technical hurdles of sustaining a dependable mobile sensing and data collection system. We revamped the StudentLife platform [342] to enhance its sensing efficiency and dependability, enabling it to support multi-year studies. This study harnesses both passive sensing data and Ecological Momentary Assessment (EMA) surveys to delve into the experiences of two undergraduate student cohorts at a four year university in the United States (starting from the year of 2017 and 2018 respectively). This extensive research spanned from September 2017 to June 2022. Consequently, this data set is an ideal fit for our objective: to evaluate and reduce
6.3 Dataset

model degradation in response to temporal environment drifts.

6.3.1 Data Collection

The study involved 215 first-year students from a four-year university in the United States, with 146 (67.8%) being female and all participants aged between 18 to 22 years at the time of enrollment. The university’s registrar informed all incoming students about the study, which started recruiting from August 2017 (cohort 2021) to November 2018 (cohort 2022). The students provided consent at the start of their first academic year. The university’s Institutional Review Board (IRB) approved the study. Initially, students had to meet the eligibility criteria which included possession of an Android or iOS smartphone compatible with the study’s mobile application. Subsequently, interested students completed an online survey to complete their eligibility requirements. Eligible students then filled out more online surveys and installed a data collection app on their smartphones. This app gathered behavioral mobile sensing data and mental health assessments through Ecological Momentary Assessments (EMA). Participants were compensated $10 per week for completing the EMA surveys but received no additional payment for the passive data collection.

6.3.2 Passive Sensing Data

Our mobile app leverages the built-in sensors of smartphones to collect data, which is then processed to derive informative features. For a complete list summary of features captured by the app is shown in Table 6.1. In addition to the features described in Chapter 1 (Section 1.2.3), we also compute several new features.
Table 6.1: Features: The table below lists the passive sensing data we collect from the participants. Daily and epoch-based sensing features are computed across meaningful daily epochs: night/morning (12am-9am), day (9am-6pm), and evening (6pm-12am), allowing analysis of behavioral features and trends across different periods of the 24-hour clock.

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Activity</td>
<td>Walking / sedentary (still) / biking / running duration</td>
</tr>
<tr>
<td></td>
<td>duration in vehicle, number of steps</td>
</tr>
<tr>
<td>Mobility &amp; Semantic</td>
<td>Distance travelled, time spent at home, workout places, study places,</td>
</tr>
<tr>
<td>Locations</td>
<td>social places, dorms, others’ dorm, greek houses, max distance from campus,</td>
</tr>
<tr>
<td></td>
<td>number of locations visited</td>
</tr>
<tr>
<td>Phone usage</td>
<td>Number of phone locks &amp; unlocks, duration of phone unlock</td>
</tr>
<tr>
<td>Audio Plays</td>
<td>Number of audio plays, duration of audio plays</td>
</tr>
<tr>
<td>Sleep</td>
<td>Sleep duration, sleep start time, sleep end time</td>
</tr>
</tbody>
</table>

Audio plays. Our application tracks active audio sessions on the phone, which includes playing any form of audio-based media such as music or videos. We calculate both the frequency of audio playbacks and the cumulative duration of these audio sessions.

6.3.3 Ground Truth

We collected weekly self-reported Patient Health Questionnaire-4 (PHQ4) data from participating students. The PHQ-4 comprises the first two questions addressing anxiety, while the last two are dedicated to depression. Notably, a total score of 3 or higher on the last two questions (the PHQ-2 subscale) is indicative of potential depression. In this study, we focus on a binary classification task to determine whether a participant has experienced depressive symptoms within a term, defined as a maximum
6.4 Analysing the Domain Drift

PHQ-2 subscale score of 3 or higher.

The PHQ-4’s primary objective is to provide an ultra-brief yet accurate assessment of the core symptoms of depression and anxiety. Both components have been independently validated as effective brief screening tools. However, it is crucial to note that an elevated PHQ-4 score does not diagnose depression but rather "an indicator for further inquiry to establish the presence or absence of a clinical disorder warranting treatment" [189]. The PHQ-4 is established as an effective indicator for the presence of depressive and anxiety disorders, but its results should be interpreted as indicative of the possibility or likelihood of these conditions, not as a definitive diagnosis. A cutoff score of 3 or above on the PHQ-2 scale shows a sensitivity of 83% and a specificity of 90% for major depressive disorder [188]. Moreover, higher PHQ-4 scores are strongly correlated with functional impairment, disability days, and healthcare utilization [189].

6.4 Analysing the Domain Drift

In this section, we address our first research question: Is there evidence of domain drift in the context of assessing depression through behavior sensing over the course of college life? We focus on two major types of drift: data drift and concept drift. Both data and concept drift can significantly impact the efficacy of a machine learning model, potentially reducing its accuracy and reliability.

Data Drift pertains to changes in the data distribution over time. This type of drift can occur due to various factors such as evolving user behaviors or external environmental changes.

Concept Drift refers to alterations in the underlying relationship between a
6.4 Analysing the Domain Drift

model's inputs (user behaviors in this case) and its outputs (depression assessments). This drift can emerge when the way behaviors relate to depression changes, possibly due to shifts in lifestyle, environmental factors, or other influences.

To understand the existence of domain drift, our analysis covers three distinct phases in the study: Pre-Covid, Covid, and the New Normal (Post-Covid). The Pre-Covid phase includes six on-campus terms spanning from 2017 to 2019, representing a baseline for student behavior under normal circumstances. The Covid phase encompasses four remote terms from 2020 to 2021, reflecting the impact of the pandemic on student lifestyles. Lastly, the New Normal phase includes three on-campus terms from 2021 to 2022, illustrating the post-pandemic adaptation period.

Each term is 10 weeks long, and we examine various student behaviors such as physical activities, audio plays, mobility patterns, on-campus semantic locations, phone usage, and sleep patterns. It is important to note that during the Covid terms, the behavior concerning on-campus semantic locations is excluded from the comparison due to the shift to remote learning environments. This comprehensive analysis aims to identify and understand the shifts in behavioral patterns corresponding to each phase and their potential implications on assessing depression.

6.4.1 Data Drift

Table 6.2 details the shifts in student behaviors across three distinct periods: Baseline (Pre-Covid on-campus terms), Covid remote terms, and New-norm return-to-campus terms. Daily behavioral sensing data are aggregated by term for each student and averaged to facilitate term comparisons. Mixed Effects Models, nested within individuals, assess the variation, with contrasts between test groups (Covid terms /
new-norm terms) and the control group (Baseline terms) designated as the fixed effect (1 for test groups, 0 for the control group). Significance levels are adjusted for multiple comparisons using the Benjamini/Hochberg False Discovery Rate (FDR) correction, with the following notations: (** p < .01, ** .01 ≤ p < .05, * .05 ≤ p ≤ .10). Values shaded in gray within the table indicate non-significant results, denoting no apparent change; they are included for informational purposes only. The ‘Baseline: Pre-Covid’ column indicates the baseline mean for daily behaviors, calculated first within each student for each term, then averaged across all students and terms. Subsequent columns quantify the behavioral deviation between the test groups and the control group. The ‘Intercept’ reflects the average predicted behavior without group distinction, slightly deviating from the ‘Baseline: Pre-Covid’ mean due to the incorporation of random effects for repeated measures in the Mixed Effects Models. The ‘Coefficient (Coef)’ column reflects the behavioral differences during test terms relative to the control group. Time epochs (0-3) demarcate various segments of the day: epoch 0 for the 24-hour average, epoch 1 for morning (12 am - 9 am), epoch 2 for daytime (9 am - 6 pm), and epoch 3 for evening (6 pm - 12 am). The units of measure (e.g., hours, minutes, kilometers) are specified alongside each behavior metric.

Substantial behavioral changes among students between the control and test group were observed. In terms of physical activities, it was noted that students spent more time in vehicles and less time in foot during Covid. This shift may be attributed to students returning home, affording them more time for driving rather than the daily routine of living in dorms and walking to classes. Another significant change was the increased sedentary duration (+1.99h/day) during Covid was reasonable given the prolonged periods spent at home. Upon entering the New Normal on campus, seden-
6.4 Analysing the Domain Drift

tary behavior and time spent on foot returned to pre-Covid levels, though there was a continued preference for using vehicles, which could be linked to students transitioning to senior years and prioritizing social interactions related to career preparation over academic pursuits.

The analysis showed a marked rise in students’ use of audio play on their phones throughout the Covid period, both on duration (+1.12h/day) and frequency (+1.58 times/day), which likely reflects students’ increased solitary time at home in the absence of in-person social interaction at school. Notably, this trend persisted and remained statistically significant even after students returned to school post-Covid, although the degree of increase had diminished.

Moving to the Mobility section, it was observed that during Covid, students traveled longer distances during the daytime but shorter distance in the mornings and evenings compared to the pre-Covid period. The amplified daytime movement is likely attributable to errands necessitating longer travel, such as grocery shopping, particularly since students were residing at home, which often demanded longer trips than the more condensed campus setting. Conversely, there was a noticeable decline in events taking place at night and early morning. Post-Covid, as students returned to campus life, these patterns readjusted to more customary levels. Furthermore, during the Covid period, students frequented fewer locations, mirroring the imposed social distancing measures. Interestingly, this reduction in location visits continued into the New Normal phase. This enduring behavior may suggest an ongoing consciousness about social distancing and a commitment to self-health protection. It could also be influenced by alterations in the college curriculum during students’ final year, potentially leading to a consolidation of their daily routines and destinations.
6.4 Analysing the Domain Drift

In terms of Semantic locations, during post-Covid, students spent less time in public areas such as food places, campus clinics, or study locations like libraries compared with pre-Covid, signaling a deliberate effort to minimize time in public spaces to avoid infection. However, there was a reduction in time spent at home and a significant increase in time spent in Greek houses, possibly suggesting an enhanced need for social interaction within their community in the final year of college.

Another significant change observed was in phone usage during Covid, with an increase in duration of phone usage (+0.73h/day) and a significant decrease in the number of phone unlocks (-19.69/day) compared to the pre-Covid period, indicating a longer usage per phone unlock. Social distancing might make students more reliant on phones for communication, replacing in-person interactions at school. However, in the New Norm period, as students returned back to school, phone usage reverted to pre-Covid levels, accompanied by a decrease in the number of phone unlocks. This shift might be attributed to increased in-person communication and social demands in senior years, reducing the frequency of checking phones.

Sleeping patterns were also affected during and after Covid. A positive change during Covid was that students had more sleep compared to the pre-Covid period. Students slept earlier (1:05 AM) and woke up later (8:31 AM) during Covid compared to pre-Covid (sleep starting at 1:23 AM and ending at 8:24 AM). However, during post-Covid, the sleeping start time (1:21 AM) returned to pre-Covid levels, while waking up later (8:48 AM) persisted, possibly indicating a sustained alteration in waking habits.
6.4 Analysing the Domain Drift

Table 6.2: Evaluation of Data Drift During Baseline, Covid, and New-Norm Terms. Mixed Effects Models, nested within individuals, assess the variation, with contrasts between test groups (Covid terms / new-norm terms) and the control group (Baseline terms) designated as the fixed effect (1 for test groups, 0 for the control group). Significance levels are adjusted for multiple comparisons using the Benjamini/Hochberg False Discovery Rate (FDR) correction, with the following notations: (** p < .01, *** .01 ≤ p < .05, * .05 ≤ p ≤ .10). Values shaded in gray within the table indicate non-significant results, denoting no apparent change; they are included for informational purposes only.

<table>
<thead>
<tr>
<th>Behavior</th>
<th>Baseline</th>
<th>vs Covid</th>
<th>vs New norm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre-Covid</td>
<td>Intercept Coef</td>
<td>Intercept Coef</td>
</tr>
<tr>
<td><strong>Physical activities</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration in vehicle (hrs) [Epoch 0]</td>
<td>0.30</td>
<td>0.31</td>
<td>0.14***</td>
</tr>
<tr>
<td>Duration in vehicle (hrs) [Epoch 2]</td>
<td>0.18</td>
<td>0.18</td>
<td>0.12***</td>
</tr>
<tr>
<td>Duration in vehicle (hrs) [Epoch 3]</td>
<td>0.09</td>
<td>0.09</td>
<td>0.02***</td>
</tr>
<tr>
<td>Duration on foot (hrs) [Epoch 0]</td>
<td>3.25</td>
<td>3.23</td>
<td>-0.99***</td>
</tr>
<tr>
<td>Duration on foot (hrs) [Epoch 1]</td>
<td>0.43</td>
<td>0.43</td>
<td>-0.18***</td>
</tr>
<tr>
<td>Duration on foot (hrs) [Epoch 2]</td>
<td>1.63</td>
<td>1.62</td>
<td>-0.42***</td>
</tr>
<tr>
<td>Duration on foot (hrs) [Epoch 3]</td>
<td>1.19</td>
<td>1.18</td>
<td>-0.40***</td>
</tr>
<tr>
<td>Sedentary duration (hrs) [Epoch 0]</td>
<td>19.97</td>
<td>20.00</td>
<td>1.99***</td>
</tr>
<tr>
<td>Sedentary duration (hrs) [Epoch 1]</td>
<td>8.44</td>
<td>8.44</td>
<td>0.21***</td>
</tr>
<tr>
<td>Sedentary duration (hrs) [Epoch 2]</td>
<td>6.97</td>
<td>6.98</td>
<td>0.36***</td>
</tr>
<tr>
<td>Sedentary duration (hrs) [Epoch 3]</td>
<td>4.57</td>
<td>4.58</td>
<td>0.43***</td>
</tr>
<tr>
<td>Sedentary duration (mins) per hour at food places</td>
<td>44.73</td>
<td>N/A</td>
<td>44.85</td>
</tr>
<tr>
<td><strong>Audio plays</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audio play duration (hrs) [Epoch 0]</td>
<td>2.62</td>
<td>2.63</td>
<td>1.12***</td>
</tr>
<tr>
<td>Audio play duration (hrs) [Epoch 1]</td>
<td>0.63</td>
<td>0.63</td>
<td>0.34***</td>
</tr>
</tbody>
</table>
### 6.4 Analysing the Domain Drift

<table>
<thead>
<tr>
<th>Audio play duration (hrs) [Epoch 2]</th>
<th>1.18</th>
<th>1.18</th>
<th>0.49***</th>
<th>1.17</th>
<th>0.35***</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio play duration (hrs) [Epoch 3]</td>
<td>0.81</td>
<td>0.82</td>
<td>0.29***</td>
<td>0.81</td>
<td>0.20***</td>
</tr>
<tr>
<td>Number of audio play period [Epoch 0]</td>
<td>5.35</td>
<td>5.35</td>
<td>1.58***</td>
<td>5.32</td>
<td>0.89***</td>
</tr>
<tr>
<td>Number of audio play period [Epoch 1]</td>
<td>1.01</td>
<td>1.02</td>
<td>0.24***</td>
<td>1.00</td>
<td>0.21***</td>
</tr>
<tr>
<td>Number of audio play period [Epoch 2]</td>
<td>2.71</td>
<td>2.71</td>
<td>1.00***</td>
<td>2.69</td>
<td>0.58***</td>
</tr>
<tr>
<td>Number of audio play period [Epoch 3]</td>
<td>1.87</td>
<td>1.87</td>
<td>0.44***</td>
<td>1.86</td>
<td>0.17***</td>
</tr>
</tbody>
</table>

#### Mobility

<table>
<thead>
<tr>
<th>Distance travelled (km) [Epoch 0]</th>
<th>13.45</th>
<th>13.72</th>
<th>1.41**</th>
<th>13.59</th>
<th>0.92</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance travelled (km) [Epoch 1]</td>
<td>2.07</td>
<td>2.09</td>
<td>-0.25*</td>
<td>2.09</td>
<td>0.11</td>
</tr>
<tr>
<td>Distance travelled (km) [Epoch 2]</td>
<td>9.05</td>
<td>9.24</td>
<td>1.68***</td>
<td>9.17</td>
<td>1.79***</td>
</tr>
<tr>
<td>Distance travelled (km) [Epoch 3]</td>
<td>4.89</td>
<td>4.91</td>
<td>-0.46*</td>
<td>4.89</td>
<td>0.21</td>
</tr>
<tr>
<td>Number of locations visited [Epoch 0]</td>
<td>5.10</td>
<td>5.11</td>
<td>-3.77***</td>
<td>5.11</td>
<td>-2.57***</td>
</tr>
<tr>
<td>Number of locations visited [Epoch 1]</td>
<td>1.63</td>
<td>1.64</td>
<td>-0.58***</td>
<td>1.64</td>
<td>-0.42***</td>
</tr>
<tr>
<td>Number of locations visited [Epoch 2]</td>
<td>3.91</td>
<td>3.91</td>
<td>-2.73***</td>
<td>3.92</td>
<td>-1.83***</td>
</tr>
<tr>
<td>Number of locations visited [Epoch 3]</td>
<td>2.72</td>
<td>2.71</td>
<td>-1.64***</td>
<td>2.71</td>
<td>-1.22***</td>
</tr>
</tbody>
</table>

#### Semantic locations

<table>
<thead>
<tr>
<th>Duration at food places (hrs)</th>
<th>1.14</th>
<th>N/A</th>
<th>N/A</th>
<th>1.14</th>
<th>-0.62***</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration greek houses (hrs)</td>
<td>0.45</td>
<td>N/A</td>
<td>N/A</td>
<td>0.47</td>
<td>1.70***</td>
</tr>
<tr>
<td>Duration at campus clinics (hrs)</td>
<td>0.03</td>
<td>N/A</td>
<td>N/A</td>
<td>0.03</td>
<td>-0.03**</td>
</tr>
<tr>
<td>Duration at home (hrs)</td>
<td>10.83</td>
<td>10.77</td>
<td>0.39</td>
<td>10.76</td>
<td>-3.95***</td>
</tr>
<tr>
<td>Duration at others’ dorms (hrs)</td>
<td>1.17</td>
<td>N/A</td>
<td>N/A</td>
<td>1.15</td>
<td>-0.63***</td>
</tr>
<tr>
<td>Duration at study places (hrs)</td>
<td>4.05</td>
<td>N/A</td>
<td>N/A</td>
<td>4.01</td>
<td>-2.93***</td>
</tr>
</tbody>
</table>

#### Phone usage

<table>
<thead>
<tr>
<th>Phone usage (hr) [Epoch 0]</th>
<th>2.99</th>
<th>2.96</th>
<th>0.73***</th>
<th>2.97</th>
<th>0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone usage (hr) [Epoch 1]</td>
<td>0.51</td>
<td>0.50</td>
<td>0.07***</td>
<td>0.50</td>
<td>0.04</td>
</tr>
<tr>
<td>Phone usage (hr) [Epoch 2]</td>
<td>1.43</td>
<td>1.42</td>
<td>0.45***</td>
<td>1.42</td>
<td>0.05</td>
</tr>
</tbody>
</table>
### 6.4 Analysing the Domain Drift

<table>
<thead>
<tr>
<th>Feature</th>
<th>Epoch 0</th>
<th>Epoch 1</th>
<th>Epoch 2</th>
<th>Epoch 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone usage (hr)</td>
<td>1.05</td>
<td>1.04</td>
<td>0.21***</td>
<td>1.04</td>
</tr>
<tr>
<td>Number of phone unlocks [Epoch 0]</td>
<td>106.46</td>
<td>104.04</td>
<td>-19.68***</td>
<td>105.64</td>
</tr>
<tr>
<td>Number of phone unlocks [Epoch 2]</td>
<td>53.05</td>
<td>51.80</td>
<td>-7.01***</td>
<td>52.63</td>
</tr>
<tr>
<td>Phone usage (mins) per hour at food places</td>
<td>N/A</td>
<td>N/A</td>
<td>8.89</td>
<td>-0.91***</td>
</tr>
<tr>
<td>Number of phone unlocks per hour at food places</td>
<td>7.95</td>
<td>N/A</td>
<td>N/A</td>
<td>7.89</td>
</tr>
<tr>
<td>Phone usage (mins) per hour at home</td>
<td>7.87</td>
<td>7.78</td>
<td>1.73***</td>
<td>7.80</td>
</tr>
<tr>
<td>Number of phone unlocks per hour at home</td>
<td>3.49</td>
<td>3.41</td>
<td>-0.12</td>
<td>3.47</td>
</tr>
<tr>
<td>Number of phone unlocks per hour at other dorms</td>
<td>7.11</td>
<td>N/A</td>
<td>N/A</td>
<td>7.07</td>
</tr>
<tr>
<td>Number of phone unlocks per hour at study places</td>
<td>5.66</td>
<td>N/A</td>
<td>N/A</td>
<td>5.60</td>
</tr>
</tbody>
</table>

**Sleep**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sleep duration (hr)</td>
<td>7.02</td>
<td>7.05</td>
<td>0.39***</td>
<td>7.07</td>
</tr>
<tr>
<td>Sleep end (AM; coef in minutes)</td>
<td>8:24</td>
<td>8:25</td>
<td>6**</td>
<td>8:25</td>
</tr>
<tr>
<td>Sleep start (AM; coef in minutes)</td>
<td>1:23</td>
<td>1:22</td>
<td>-17***</td>
<td>1:21</td>
</tr>
</tbody>
</table>

### 6.4.2 Concept Drift

Table 6.3 details the shifts in the association between student behaviors and depression across three distinct periods: Baseline (Pre-Covid on-campus terms), Covid remote terms, and New-norm return-to-campus terms. The student population was
6.4 Analysing the Domain Drift

categorized into two groups: the non-depressed as the control group and the de-
pressed as the testing group. Mixed Effects Models, nested within individuals, assess
the variation, with contrasts between the test group (students with depression) and
the control group (students without depression) designated as the fixed effect (1 for
test groups, 0 for the control group). Columns quantify the behavioral deviation
between the test group and the control group. The ‘Intercept’ reflects the average
predicted behavior without group distinction. The ‘Coefficient (Coef)’ column re-
flects the behavioral differences between the test group (depressed students) relative
to the control group (non-depressed students).

In terms of physical activities, students with depression exhibited more sedentary
behavior during evenings compared to their non-depressed counterparts. However,
during Covid, this difference ceased to be significant, and the pattern persisted into
the New-norm terms. Another noteworthy change was the decreased time spent on
foot by students with depression during Covid, reverting to normal in the New-norm
terms possibly because both groups returned to school and engaged in regular walking
to classrooms.

Turning to Audio plays, during the baseline terms, the depression group invested
more time (0.23h/day). However, in the Covid terms, although students with depres-
sion still demonstrated a higher inclination towards music during the early morning
hours, the difference was no longer significant. In the new-norm terms, both groups
exhibited no significant disparity in audio play.

Regarding the Mobility factor, the depression group visited fewer locations (-
0.39/day) compared to the non-depression group students pre-Covid. However, dur-
ing and after Covid, there were no significant differences between the two groups,
6.4 Analysing the Domain Drift

suggesting a shared constraint on mobility during the Covid period, which persisted into the new-norm terms.

In terms of phone usage during baseline terms, the depression group exhibited higher phone usage, particularly in Epoch 0 (24-hour average) and Epoch 1 (12 am-9 am). Notably, the depression group demonstrated more significantly increased phone usage, also particularly during 12 am - 9 am, during the Covid period, underscoring the reduced sleep. However after Covid, the phone usage patterns in both groups did not show significant differences.

During the baseline terms, individuals in the depression group showed higher phone usage, particularly noticeable in Epoch 0 (24-hour average) and Epoch 1 (12 am to 9 am). This trend was even more pronounced during the Covid period, especially in the 12 am to 9 am window. However, post-Covid, the phone usage patterns between the depressive and non-depressive groups did not exhibit significant differences. An additional interesting finding is that during both the baseline and Covid terms, the depressive group had higher phone usage at home. Yet, in the post-Covid period, their phone usage at home decreased relative to the non-depressive group. This concept drift is somewhat counterintuitive and may be attributed to certain specific factors in the new norm that led to an overall increase in phone usage at home, as suggested by the data presented in Table 6.2. It appears that the non-depressive group experienced an even greater increase in phone usage at home post-Covid, thereby diminishing the previously observed differences in home phone usage between the two groups.

During the baseline terms, students in the depression group had less sleep compared to their non-depressed counterparts. However, during and after Covid, although the depression group still experienced less sleep, the difference between the two groups
### 6.4 Analysing the Domain Drift

was not significant.

Table 6.3: Evaluation of Concept Drift (the association between behavior and depression) during Baseline, Covid, and New-Norm Terms. The student population was categorized into two groups: the non-depressed as the control group and the depressed as the testing group. Mixed Effects Models, nested within individuals, assess the variation, with contrasts between the test group (students with depression) and the control group (students without depression) designated as the fixed effect (1 for test groups, 0 for the control group). Significance levels are adjusted for multiple comparisons using the Benjamini/Hochberg False Discovery Rate (FDR) correction, with the following notations: (** p < .01, * .01 ≤ p < .05, * .05 ≤ p ≤ .10). Values that are not shaded gray and lack notations had p < .05 before FDR but do not maintain significance after the FDR correction was applied. Values shaded in gray within the table indicate non-significant results, denoting no apparent change; they are included for informational purposes only.

<table>
<thead>
<tr>
<th>Daily Behavior</th>
<th>Baseline terms</th>
<th>Covid terms</th>
<th>New-norm terms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-dep vs Dep</td>
<td>Non-dep vs Dep</td>
<td>Non-dep vs Dep</td>
</tr>
<tr>
<td></td>
<td>Intercept</td>
<td>Coef</td>
<td>Intercept</td>
</tr>
<tr>
<td><strong>Physical activities</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sedentary duration (hrs) [Epoch 3]</td>
<td>4.54</td>
<td>0.07</td>
<td>5.01</td>
</tr>
<tr>
<td>Duration on foot (hrs) [Epoch 0]</td>
<td>3.29</td>
<td>-0.10</td>
<td>2.22</td>
</tr>
<tr>
<td>Duration on foot (hrs) [Epoch 2]</td>
<td>1.64</td>
<td>-0.03</td>
<td>1.20</td>
</tr>
<tr>
<td><strong>Audio plays</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audio play duration (hrs) [Epoch 0]</td>
<td>2.57</td>
<td>0.23</td>
<td>3.60</td>
</tr>
<tr>
<td>Audio play duration (hrs) [Epoch 2]</td>
<td>1.15</td>
<td>0.10</td>
<td>1.57</td>
</tr>
<tr>
<td>Number of audio play periods [Epoch 0]</td>
<td>5.25</td>
<td>0.35</td>
<td>6.77</td>
</tr>
<tr>
<td>Number of audio play periods [Epoch 1]</td>
<td>0.97</td>
<td>0.14**</td>
<td>1.21</td>
</tr>
<tr>
<td>Number of audio play periods [Epoch 2]</td>
<td>2.66</td>
<td>0.17</td>
<td>3.60</td>
</tr>
<tr>
<td><strong>Mobility</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of location visited [Epoch 0]</td>
<td>5.24</td>
<td>-0.39**</td>
<td>1.35</td>
</tr>
<tr>
<td>Number of location visited [Epoch 1]</td>
<td>1.68</td>
<td>-0.13**</td>
<td>1.05</td>
</tr>
<tr>
<td>Number of location visited [Epoch 2]</td>
<td>4.00</td>
<td>-0.28*</td>
<td>1.20</td>
</tr>
<tr>
<td>Number of location visited [Epoch 3]</td>
<td>2.76</td>
<td>-0.15</td>
<td>1.08</td>
</tr>
</tbody>
</table>

209
6.5 Detecting Model Degradation

<table>
<thead>
<tr>
<th>Semantic locations</th>
<th></th>
<th></th>
<th>N/A</th>
<th>N/A</th>
<th>0.002</th>
<th>0.003</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration at on-campus clinics (hours)</td>
<td>0.02</td>
<td>0.05**</td>
<td>N/A</td>
<td>N/A</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td>Duration at workout places (hours)</td>
<td>0.37</td>
<td>-0.09</td>
<td>N/A</td>
<td>N/A</td>
<td>0.15</td>
<td>-0.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phone usage</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone usage (hours) [Epoch 0]</td>
<td>2.91</td>
<td>0.17</td>
<td>3.21</td>
<td>0.36*</td>
<td>3.36</td>
<td>0.01</td>
</tr>
<tr>
<td>Phone usage (hours) [Epoch 1]</td>
<td>0.49</td>
<td>0.05</td>
<td>0.54</td>
<td>0.10*</td>
<td>0.61</td>
<td>0.05</td>
</tr>
<tr>
<td>Phone usage (hours) [Epoch 2]</td>
<td>1.40</td>
<td>0.07</td>
<td>1.82</td>
<td>0.15</td>
<td>1.59</td>
<td>0.03</td>
</tr>
<tr>
<td>Phone usage (hours) [Epoch 3]</td>
<td>1.03</td>
<td>0.04</td>
<td>1.23</td>
<td>0.11</td>
<td>1.15</td>
<td>-0.05</td>
</tr>
<tr>
<td>Phone usage (minutes) per hour at home</td>
<td>7.62</td>
<td>0.61</td>
<td>9.19</td>
<td>0.65</td>
<td>9.50</td>
<td>-1.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sleep</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sleep duration (hours)</td>
<td>7.12</td>
<td>-0.26</td>
<td>7.58</td>
<td>-0.14</td>
<td>7.54</td>
<td>-0.39</td>
</tr>
</tbody>
</table>

6.5 Detecting Model Degradation

In this section, we discuss our second research question: *Does the domain drift cause significant model degradation?* To answer this question, we divide the data into a source dataset and a few target dataset. The terms included in each dataset are specified in Table 6.4. We include 6 terms from 2017 fall to 2019 spring so that every student has at least one year on-campus data covering terms in different seasons in the source dataset (the 2nd cohort started from 2018 fall). The domain drift may happen in various events during the college lives, including (but not restricted to) studying on campus, leaving campus during holidays, remote classes during COVID, back to in-person classes and the new normal. To get a more comprehensive understanding of the model degradation at different phases of campus life on and off campus, we include more scenarios such as winter and summer holidays than what has discussed
6.5 Detecting Model Degradation

in Section 6.4. The target COVID and target new-norm dataset are selected based on the college remote term and return-to-campus policy during COVID.

Table 6.4: Source and target dataset

<table>
<thead>
<tr>
<th>Category</th>
<th>Term/Holiday Period</th>
<th>Length of Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source terms</td>
<td>2017 fall term (Sep-Nov)</td>
<td>10 weeks</td>
</tr>
<tr>
<td></td>
<td>2018 winter term (Jan-Mar)</td>
<td>10 weeks</td>
</tr>
<tr>
<td></td>
<td>2018 spring term (Mar-Jun)</td>
<td>10 weeks</td>
</tr>
<tr>
<td></td>
<td>2018 fall term (Sep-Nov)</td>
<td>10 weeks</td>
</tr>
<tr>
<td></td>
<td>2019 winter term (Jan-Mar)</td>
<td>10 weeks</td>
</tr>
<tr>
<td></td>
<td>2019 spring term (Mar-Jun)</td>
<td>10 weeks</td>
</tr>
<tr>
<td>Target</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Target normal terms</td>
<td>2019 fall term (Sep-Nov)</td>
<td>10 weeks</td>
</tr>
<tr>
<td>Target winter holidays</td>
<td>2017 winter holiday (Nov-Dec)</td>
<td>5 weeks</td>
</tr>
<tr>
<td></td>
<td>2018 winter holiday (Nov-Dec)</td>
<td>5 weeks</td>
</tr>
<tr>
<td></td>
<td>2019 winter holiday (Nov-Dec)</td>
<td>5 weeks</td>
</tr>
<tr>
<td>Target summer holidays</td>
<td>2018 summer holiday (Jun-Sep)</td>
<td>14 weeks</td>
</tr>
<tr>
<td></td>
<td>2019 summer holiday (Jun-Sep)</td>
<td>14 weeks</td>
</tr>
<tr>
<td>Target COVID off campus terms</td>
<td>2020 spring term (Mar-Jun)</td>
<td>10 weeks</td>
</tr>
<tr>
<td></td>
<td>2020 fall term (Sep-Nov)</td>
<td>10 weeks</td>
</tr>
<tr>
<td></td>
<td>2021 winter term (Jan-Mar)</td>
<td>10 weeks</td>
</tr>
<tr>
<td></td>
<td>2021 spring term (Mar-Jun)</td>
<td>10 weeks</td>
</tr>
<tr>
<td>Target new-norm terms</td>
<td>2021 fall term (Sep-Nov)</td>
<td>10 weeks</td>
</tr>
<tr>
<td></td>
<td>2022 winter term (Jan-Mar)</td>
<td>10 weeks</td>
</tr>
<tr>
<td></td>
<td>2022 spring term (Mar-Jun)</td>
<td>10 weeks</td>
</tr>
</tbody>
</table>

We focus on a binary classification task to determine whether a participant has ever experienced depressive symptoms within a term, defined as a maximum PHQ-2 subscale score of 3 or higher. When aggregating a student’s weekly self-reports over a term, we adjust the weight of samples as outlined in Table 6.5. This approach aims to mitigate some labeling ambiguity and focuses the model on effectively identifying individuals with severe and persistent depressive symptoms, rather than briefly
6.5 Detecting Model Degradation

occurring or mild symptoms that rapidly resolve. This methodology reflects an intention to enhance the model’s utility in detecting more clinically significant cases of depression.

Table 6.5: Data-point weight adjustment based on PHQ-2 scores and symptom lasting length when aggregating weekly self reports to a term

<table>
<thead>
<tr>
<th>Maximum PHQ-2 score in term</th>
<th>Weeks with PHQ-2 score ( \geq 3 ) in term</th>
<th>Data-point weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>( \geq 3 )</td>
<td>1</td>
<td>0 (remove)</td>
</tr>
<tr>
<td>( \geq 3 )</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>( \geq 3 )</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>( \geq 3 )</td>
<td>( \geq 3 )</td>
<td>3</td>
</tr>
</tbody>
</table>

6.5.1 Training Details

We first developed a model using the source dataset, where the data was randomly divided into training, validation, and testing sets in a 60%, 20%, and 20% ratio, respectively. Each data point corresponds to a student in a term, with 101 daily average sensed behavior features and a depression label derived from the highest self-reported PHQ-4 score for the term, weighted according to frequency as detailed in Section 3.1. We standardized features in the training and validation sets of the source dataset, applying the same standardization function to the testing data and all target datasets. We build a fully connected deep neural network, initially passing through a batch normalization layer with ReLU activation, followed by four fully connected hidden layers. Each layer had [256, 128, 64, 32] hidden units, respectively,
6.5 Detecting Model Degradation

each followed by a ReLU activation layer and a dropout layer with a rate of 0.5. The output size was set to 1, and we used BCEWithLogitsLoss as the loss function. We used Adam as the optimizer with a learning rate of 0.001. After evaluating the model’s performance on the training and validation sets, we determined that the optimal training duration was 80 epochs. We then assessed the trained model on the testing set of the source dataset and on each whole target dataset. This process was repeated 10 times with different initial random weight settings, resulting in 10 trained models. We evaluated these models and reported the mean and standard deviation of their performance on both the testing and target datasets. Our evaluation included two performance metrics, which are also often used in other model generalizability studies [367]: (1) ROC-AUC, which is an overall result across varying decision boundaries in a detection problem; and (2) balanced accuracy, which is the mean of sensitivity (true positive rate) and specificity (true negative rate) and is robust to class imbalance.

6.5.2 Result Analysis

Figure 6.1 displays a bar plot of the model performances, revealing several observations. Firstly, the model demonstrates effectiveness on the source dataset’s testing set, achieving a ROC AUC of 0.82 and a balanced accuracy of 0.75. This result underscores the model’s capability to predict students’ depression status accurately in the absence of domain drift. Secondly, a notable reduction in model performance is observed across all target datasets, signifying the impact of domain drift on model degradation. Thirdly, the extent of this degradation varies among different target datasets, depending on the degree of domain drift. For on-campus terms, such as the target-normal-terms (ROC AUC 0.65 and balanced accuracy 0.61) and target-new-
6.5 Detecting Model Degradation

norm-terms (ROC AUC 0.67 and balanced accuracy 0.63), the model still outperforms its performance during off-campus holidays, including target-winter-holidays (ROC AUC 0.60 and balanced accuracy 0.56) and target-summer-holidays (ROC AUC 0.61 and balanced accuracy 0.56). The poorest performance is observed during the COVID-remote-terms, with a ROC AUC of 0.57 and balanced accuracy of 0.55.

Figure 6.1: Model performance degradation of predicting depression in term when applying a model trained on source data (orange) on various target data (blue)
6.6 Mitigating Model Degradation using Domain Adaptation Algorithms

In this section, we address our third research question: *Can we use adaptation techniques to counteract this model degradation over time?*

6.6.1 Unsupervised Domain Adaptation

In unsupervised domain adaptation, a model is trained on labeled data from a source domain and then moved to a target domain without any labeled data in the target domain [258, 275, 215]. We implemented and evaluated two prominent domain adaptation algorithms as follows, both of which offer unique approaches to bridging the divide between source and target domains:

(a) DANN (Domain-Adversarial Neural Network) [120]: a representation learning technique that adversarially trains the generator and discriminator. The discriminator is trained to distinguish different domains, while the generator is trained to fool the discriminator to learn domain-invariant feature representations.

(b) MCD (Maximum Classifier Discrepancy) [291]: a representation learning technique that attempts to align distributions of source and target by utilizing the task-specific decision boundaries. It maximizes the discrepancy between two classifiers’ outputs to detect target samples that are far from the support of the source. A feature generator learns to generate target features near the support to minimize the discrepancy.
6.6 Mitigating Model Degradation using Domain Adaptation Algorithms

We chose DANN for its ability to learn feature representations that are not only discriminative for the learning task but also indistinguishable in terms of the domain origin, by introducing an adversarial domain classifier. This method is particularly well-suited for scenarios where the main challenge lies in aligning the feature distribution of the source and target domains without requiring alignment of label distributions. We also chose MCD for its innovative use of two classifiers to maximize the discrepancy on the target domain, encouraging the model to learn more generalizable features by enforcing that the two classifiers disagree with each other as much as possible on the target samples. This approach is especially compelling when we suspect that the target domain has ambiguous or noisy labels. By employing both DANN and MCD, we aim to cover a broader spectrum of domain adaptation challenges, assessing which method better facilitates the adaptation of models in the presence of varying degrees of domain shift for accessing depression of students.

In the unsupervised domain adaptation approach, we utilized the features and depression labels of the entire source-terms dataset, along with the features of the target dataset, for training purposes. An additional aspect of this training involved the use of domain labels, which indicate whether the data originated from the source or target dataset. It’s important to note that the depression labels from the target dataset were exclusively reserved for testing; they were not employed in the training phase. The objective is to enable the model to learn and adapt to the differences between the source and target datasets, inferred from the domain labels, thereby improving its performance on the target dataset without the necessity of labeled data from this target domain. The true labels of the target dataset are then used post-training to evaluate the model’s performance in adapting to and predicting in the
6.6 Mitigating Model Degradation using Domain Adaptation Algorithms

new domain.

6.6.2 Result Analysis

Figure 6.2 displays a bar plot illustrating the model performances with domain adaptation, yielding several interesting insights.

![Bar plot showing model performance with domain adaptation](image)

Figure 6.2: Model performance of using DANN (blue) and MCD (green) domain adaptation algorithms on target dataset. The gray bar is the performance without using any adaptation algorithms and is included for comparison purpose.

Firstly, domain adaptation proves effective only for Pre-Covid target-normal-terms...
and target-Covid-remote terms. Specifically, DANN (Domain-Adversarial Neural Network) enhances the ROC-AUC from 0.65 to 0.71 and the balanced accuracy from 0.61 to 0.65 for Pre-Covid target-normal-terms. In the case of target-Covid-remote terms, DANN improves the ROC-AUC from 0.57 to 0.61 and the balanced accuracy from 0.55 to 0.58. However, no significant improvements are observed for the target-winter-holidays and target-summer-holidays. This may be attributed to the inability of domain adaptation to address scenarios where the underlying logic and concepts of the data undergo substantial changes. For example, during academic terms, daily behaviors are largely influenced by class schedules, and stress from coursework and exams is a major factor in depression. This pattern persists even during remote terms in the Covid era. However, holiday behaviors deviate from this academic rhythm, leading to limited gains from applying source term learning to holiday periods. Interestingly, after domain adaptation, performance on target-Covid-remote-terms surpasses that of target-holidays, despite initially being poorer. This suggests that although data distributions differ markedly between on-campus and off-campus settings, domain adaptation may still identify domain-independent features, facilitating the transfer of knowledge from on-campus to off-campus term scenarios.

Secondly, DANN seems to be more effective than MCD, particularly in terms of the ROC-AUC metric. This observation suggests that the model degradation in depression assessment across different eras of college life is more likely due to changes in feature distribution rather than label distribution.

Thirdly, it is noteworthy that domain adaptation does not improve performance on the target-new-norm-terms. In fact, after applying domain adaptation, the performance on these terms is inferior compared to the target-normal-terms, with a
6.6 Mitigating Model Degradation using Domain Adaptation Algorithms

ROC-AUC of 0.69 versus 0.71 and a balanced accuracy of 0.61 versus 0.65. This difference in performance could be attributed to a larger concept drift in the post-Covid new norm, as detailed in Table 6.3. Additionally, the target-new-norm-terms are temporally further from the source terms compared to the target-normal-terms.

6.6.3 Semi-supervised Domain Adaptation with Anchor Data

The findings in the analysis above underscore a critical challenge in domain adaptation: even though some improvement is observed when applying domain adaptation techniques, achieving model performance on the target dataset comparable to that on the source dataset remains a significant hurdle, particularly in the absence of labels in the target dataset. Namely, it suggests intrinsic limitations of unsupervised domain adaptation approaches in scenarios where domain shift is significant and complex.

This points towards the potential benefits of semi-supervised approaches, namely, the necessity of incorporating some form of "anchor data" in the target dataset. Anchor data could be a subset of the target dataset with known labels, which would provide a reference point for the model to better understand and adapt to the new domain. This anchor data acts as a bridge, helping the model to recalibrate and adjust its parameters more effectively for the target domain. The presence of such data can be crucial in fine-tuning the model to the nuances and specific characteristics of the target dataset, thereby enhancing its accuracy and reliability.

In the previous subsection, we observed some improvement in the target-Covid-remote-terms through unsupervised domain adaptation, yet the performance still lagged behind that of the on-campus terms. In addition, the return-to-campus target-new-norm-terms did not exhibit similar improvements, possibly due to their greater
Mitigating Model Degradation using Domain Adaptation Algorithms

temporal distance from the source dataset. This led us to hypothesize that semi-supervised domain adaptation, incorporating anchor data, could potentially enhance performance. In this subsection, we explore semi-supervised domain adaptation on these two specific target datasets, each comprising 4 and 3 terms respectively, as outlined in Table 6.4. For each dataset, we allowed access to the depression labels of the first term, incorporating these labels into the loss function to optimize during the training process. This approach represents a semi-supervised methodology, where a limited amount of labeled data from the target domain (the first term in each dataset) is used in conjunction with the larger unlabeled portion of the target dataset.

We then evaluated the model’s performance on the remaining terms within each dataset, treating the first term as anchor data to guide the adaptation process. The rationale behind this approach is that the anchor data can provide valuable insights into the target domain, allowing the model to adjust more effectively to the new context. The results of this semi-supervised domain adaptation, including comparisons with previous unsupervised approaches and the resulting performance improvements, are presented in Figure 6.3.

The application of anchor labels in semi-supervised domain adaptation yielded notable improvements in model performance, particularly when compared to the results obtained from unsupervised domain adaptation. For the target-Covid-remote-terms, the introduction of anchor labels enhanced the model’s ROC-AUC from 0.61 to 0.63 and the balanced accuracy from 0.58 to 0.61. This improvement, while modest, indicates the effectiveness of integrating some labeled data in the adaptation process.

More significantly, the impact of anchor labels on the target-new-norm-terms was quite substantial. The ROC-AUC increased from 0.68 to 0.76, and the balanced accu-
6.7 Discussion

Accuracy saw a remarkable rise from 0.62 to 0.71. This performance is notably close to the balanced accuracy of 0.75 achieved on the source dataset, underscoring the potential of semi-supervised domain adaptation in bridging the gap between the source and target domains.

Figure 6.3: Model performance of using semi-supervised DANN (blue) and MCD (green) domain adaptation algorithms with anchor labels on target dataset.

6.7 Discussion

6.7.1 Summary and Findings

In our experimental analysis, we make several key findings. Initially, we employed mixed-effect models to understand domain drift from a temporal perspective. We observed significant data drift across baseline (pre-covid), covid, and new norm time
6.7 Discussion

periods. In particular, features such as sedentary duration, audio play duration, travel distance, phone duration, and sleeping patterns differed across the time periods. For example, individuals used their phones for longer periods of time during covid than other terms. On another note, we observed that students slept longer and went to bed earlier during covid, and this effect continued after covid in the new norm. After statistically observing distribution shifts across term periods, we sought to understand model degradation with the task of depression detection. Thus, we trained a deep neural network with one source dataset composed of six terms and many target datasets with different terms. Our results indicate the models degrade across all target datasets to varying degrees (section 6.5.2). It is unsurprising to observe greater degradation in model performance during holidays and remote terms compared to on-campus terms. Notably, the least degradation was seen in the new-norm terms, which may imply that these terms are closer to the baseline data than other target settings. The lesser degradation in the new-norm terms compared to the target-normal-terms could be attributed to the university’s policy allowing students to participate in off-campus programs during their 2nd and 3rd academic years, whereas they are required to be on-campus in their final year.

To reduce the effect of model degradation, we studied unsupervised domain adaptation (UDA) and semi-supervised domain adaptation (SSDA). While UDA approaches improved performance of normal and covid-remote terms, it did not perform well for holiday terms. This may be attributed the significant data and concept shift during holidays. For example, the student’s schedule at home or internships will be much different that during the academic term. Consequently, the scarcity of domain-invariant feature representations undermines the model’s effectiveness. Additionally, the lack of
improvement in performance on target-new-norm terms through UDA suggests that large temporal shifts (approx. 2 years), may pose challenges to adaptation without access to new mental health labels in the post-Covid new norm.

Furthermore, DANN performed better than MCD indicating that the degradation is more likely due to the behavior distribution shift instead of ambiguous or noisy labels. The SSDA framework using limited target data yielded better performance improvements over UDA. Notably, the use of anchor data helps guide the training process better. Furthermore, previous studies have shown that SSDA approaches perform well with small datasets [269].

6.7.2 Implications

In our study, we explore the impact of data distribution changes on model efficacy over a span of four years. Our finds have significant implications across various domains such as personalization and continual learning strategies. For instance, a notable degradation in model performance at an individual level over time emphasizes the necessity for developing personalized models that can adapt to changes in individual behaviors and preferences. This finding indicates that models trained on historical data degrade rapidly. Consequently, incorporating adaptive learning mechanisms that can update or edit models continuously as new streams of data arrive is crucial. Doing so will not only enhance the user experience by providing more accurate and relevant outputs but also contributes to the advancement of continual learning systems that can evolve and improve over time.
Chapter 7

Conclusion

This thesis has tackled various technical challenges within the realm of mobile sensing. Our efforts have broadened the scope of mobile sensing capabilities to accommodate diverse and complex environments. By developing power-efficient smartphone sensing systems, we advocate for expansive, longitudinal research spanning various demographics. We aim for these technologies to be more flexible, inclusive, and proficient in documenting the intricate dynamics of human existence, thus providing a deeper understanding in the realm of mental health and beyond.

The contribution of this thesis are summarized as follows:

- Firstly, our exploration into personality traits through mobile sensing highlights the intricate dance of within-person variability and its capacity to predict personality dimensions such as extraversion and agreeableness. This novel approach offers a scalable and passive alternative for personality assessment. The implications extend beyond academia, suggesting enhancements to recommendation systems, recruitment processes, and broader human-centered applications.
Conclusion

- Secondly, the investigation into social functioning among individuals with mental health disorders like schizophrenia demonstrates mobile sensing’s potential to transcend subjective evaluations. By correlating mobile sensing data with detailed measures of social functioning, we unlock new pathways for understanding and intervening in mental health and aging, with precision previously unattainable.

- Thirdly, the study on auditory verbal hallucinations (AVH) showcases the feasibility of employing voice diary EMAs alongside mobile sensing for assessing AVH severity. This approach not only elevates the accuracy of mental health evaluations but also underscores the viability of technology-driven self-tracking tools, setting the stage for future mental health assessments.

- Fourthly, our work with first-generation college students underscores the unique challenges and mental health risks they face, leveraging mobile sensing to unearth behavioral patterns linked to depression and anxiety. This study not only provides a foundation for more nuanced mental health predictions but also promises longitudinal insights into the college journey of these students, emphasizing the need for tailored support mechanisms.

- Lastly, the examination of model degradation over four years stresses the urgency of developing adaptive models to combat the diminishing effectiveness of static predictive systems. This finding urges the adoption of continual learning strategies, ensuring that models remain relevant over time, thus enhancing the accuracy and utility of mobile sensing applications.
7.1 Insights

In our AVH study, we pioneered the use of Google ads and online recruiting methods on a large scale targeting a sensitive group, demonstrating their effectiveness in significantly reducing recruitment costs and enhancing participant diversity. Throughout the recruitment process, we encountered challenges such as attempts by some individuals to exploit the system by submitting duplicate applications for multiple compensations. We identified several gaming patterns in the early months of the study and continuously improved our system to mitigate these issues. To prevent fraudulent applications, we implemented checks where potential gamers’ applications were interrupted, requiring them to verify their information via a phone call before proceeding. Some checks can only be done after the app was installed; in this case we retained the right to exclude potential gamers from the study. Identified gamer patterns included: (1) use of a similar email address to one previously registered, (2) accessing our study webpage by directly entering the URL rather than being redirected from a Google ads link, (3) inconsistencies between the claimed state of residence and the IP address location, (4) reinstallation of the app on a device previously used in the study, (5) GPS data collected in the week following app installation that showed location patterns similar to those of an existing user. By successfully identifying and addressing these gaming patterns, we ensured the integrity and validity of our dataset. The insights gained from these experiences are invaluable and can undoubtedly benefit similar studies in the future, particularly those involving sensitive groups or large-scale online recruitment.

Our studies have significantly advanced the field of mental health sensing, yet we
7.1 Insights

see persistent challenges in this area. Researchers collect data from different populations individually, employing similar processes and settings. However, it is evident that different study populations can significantly impact the results. Consequently, findings in this field often suffer from overfitting to specific environments and lack generalizability. While certain correlations, such as between mobility and mental health, may be consistent across various studies, adapting machine learning approaches to new problem sets typically requires considerable additional work, including the incorporation of new features and the design of specific models.

This limitation hinders the broader adoption of mobile sensing in real-world applications. One primary reason for these varying results is the gap between the behavior patterns captured through current sensing methods and the entirety of human life. Mobile phone-based sensing, for instance, misses significant information when the user is away from their device and struggles to capture fine-grained activities, especially indoor human activities. Such gaps likely contribute to the observed variability in study outcomes. Assuming that mental health can be inferred from behavior, current methodologies in this field capture only a fraction of the complete data. To bridge this gap, the future of mental health sensing must incorporate additional peripheral sensing tools that can provide a more comprehensive view of daily human behavior. These enhancements are crucial for developing more accurate and generalizable mental health assessment tools, driving forward the field’s capability to deliver real-world impact.

I remain optimistic about the future of mobile sensing, especially with the emergence of Large Language Models (LLMs). Our work with audio diaries has demonstrated that even a modest addition of speech data can be highly beneficial. I an-
ticipate that LLM-based AI assistants will soon be integrated into mobile sensing tools, significantly enhancing the quality of data collected from participants. On-device LLMs could notably improve the richness of data by enabling more complex and nuanced data capture directly from users. Moreover, LLM assistants could play a crucial role when there are gaps in sensing data, such as when a phone is not carried by the user. In these instances, the LLM assistant could engage with the user through a quick chat to gather essential ground truth data, ensuring that the sensing dataset remains robust and comprehensive.

The integration of AI in mobile sensing is poised to substantially enhance the quality of sensed behavioral data. This advancement is likely to bolster confidence in mobile sensing technologies, facilitating the creation of shared datasets and feature sets that are more generalizable than those currently available. Looking ahead, we could move beyond merely analyzing correlations and machine learning-based measurements to exploring causal relationships between behavior and mental health. This deeper understanding is crucial for developing personalized intervention plans that improve mental health and overall well-being.

7.2 Future Work

There are many ways our work can be improved and extended. There is a lot of wide-open areas that have not been fully explored yet. Future studies could work on the following areas.

Enhancing Data Privacy and Security. Future work could explore more advanced data privacy techniques, such as secure multi-party computation, to protect participant data while still extracting meaningful insights. This would address grow-
7.2 Future Work

ing concerns about data security and privacy in mobile sensing research.

**Longitudinal Studies on a Larger Scale.** Conducting longer-term studies with a more diverse and larger population would help validate the robustness of the findings and allow for the exploration of more complex behavioral patterns over time.

**Integration of New Wearable Devices.** Beyond smartphones, incorporating data from new wearable devices such as smart glasses, smart watches, and smart rings can provide more comprehensive insights into physiological and health-related parameters. Also, AR and VR technologies can be used in future studies to simulate environments for controlled behavioral experiments. They offer new ways to study social interactions, cognitive behaviors, and stress responses in immersive, customizable settings. Future studies could explore how these devices to offer a fuller picture of an individual’s mental and physical health.

**Leveraging Large Language Models (LLMs).** The use of LLMs can revolutionize the analysis of text and audio data collected in mobile sensing studies. These models can help in extracting deeper insights from voice diaries, textual inputs, and social media interactions, potentially offering novel indicators of mental health status or personality traits.

**Advanced Machine Learning and Deep Learning Approaches.** Exploring new advancements in machine learning and deep learning, such as transformers and graph neural networks, could enhance the analysis of temporal and spatial data patterns. This would improve the prediction of mental health issues and social behaviors from mobile sensing data.
7.3 Final Comment

**Personalization of Interventions.** Recognizing that individuals respond differently to various interventions, systems should be designed to personalize the response based on the user’s history, preferences, and specific needs. This might involve machine learning algorithms that adapt over time. It’s also important to establish validated thresholds for these behaviors and symptoms that, when exceeded, indicate a need for intervention. These thresholds should be based on empirical research and clinical expertise.

7.3 Final Comment

The exploration and advancements in mobile sensing, along with the integration of new technologies and ethical intervention constructs, highlight a promising frontier in understanding and improving human mental health and behavior. As we move forward, it’s imperative that we continue to prioritize ethical considerations, privacy, and personalization to ensure that these technologies not only advance our capabilities but also respect the individuals they are designed to serve. The journey ahead is both exciting and challenging, offering endless possibilities for innovation and impact in the realm of mental health and beyond.
Chapter 8

Refereed Publications as a Ph.D. Candidate

My refereed publications as a Ph.D. candidate are listed below. Work in preparation and technical reports are omitted.

Conference/Workshop Publications


Rui Wang, Weichen Wang, Mikio Obuchi, Emily Scherer, Rachel Brian, Dror Ben-Zeev, Tanzeem Choudhury, John Kane, Martar Hauser, Megan Walsh, and Andrew Campbell. “On Predicting Relapse in Schizophrenia using Mobile Sensing in a Randomized Control Trial”. In: 2020 IEEE International Conference on Pervasive
Refereed Publications as a Ph.D. Candidate

Computing and Communications (PerCom). 2020, pp. 1–8. doi: 10.1109/PerCom45495.2020.9127365


Refereed Publications as a Ph.D. Candidate


Journal Publications


Rui Wang, Weichen Wang, Min Hane Aung, Dror Ben-Zeev, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A. Scherer,
Refereed Publications as a Ph.D. Candidate


Refereed Publications as a Ph.D. Candidate


Sandrine R. Müller, Heinrich Peters, Sandra C. Matz, Weichen Wang, and Gabriella M. Harari. “Investigating the Relationships between Mobility Behaviours and Indicators of Subjective Well-Being Using Smartphone-Based Experience Sampling and
Refereed Publications as a Ph.D. Candidate


Dror Ben-Zeev, Benjamin Buck, Ayesha Chander, Rachel Brian, Weichen Wang, David Atkins, Carolyn J Brenner, Trevor Cohen, Andrew Campbell, and Jeffrey Munson. “Mobile RDoC: Using Smartphones to Understand the Relationship Between Auditory Verbal Hallucinations and Need for Care”. In: Schizophrenia Bul-
Refereed Publications as a Ph.D. Candidate


237
Refereed Publications as a Ph.D. Candidate

In: JMIR Form Res 5.6 (June 2021), e23118. issn: 2561-326X. doi: 10.2196/23118.


Refereed Publications as a Ph.D. Candidate


Refereed Publications as a Ph.D. Candidate


Book Chapter

Bibliography


[24] Dror Ben-Zeev, Rachel Brian, Rui Wang, Weichen Wang, Andrew T Campbell, Min SH Aung, Michael Merrill, Vincent WS Tseng, Tanzeem Choudhury,
BIBLIOGRAPHY


[34] Josef Bless, Runar Smelror, Ingrid Agartz, and Kenneth Hugdahl, *Sa110. using a smartphone app to assess auditory hallucinations in adolescent schizophrenia: Is this the way to go for better control over voices?*, Schizophrenia bulletin 43 (2017), no. Suppl 1, S152.


[37] Mehdi Boukhechba, Alexander R. Daros, Karl Fua, Philip I. Chow, Bethany A. Teachman, and Laura E. Barnes, *DemonicSalmon: Monitoring mental health*
and social interactions of college students using smartphones, Smart Health (2018).


[40] Noli Brazil and Matthew Andersson, Mental Well-Being and Changes in Peer Ability From High School to College, Youth & Society (2018), 0044118X1876452 (en).


[62] Xingui Chen, Gong-Jun Ji, Chunyan Zhu, Xiaomeng Bai, Lu Wang, Kongliang He, Yaxiang Gao, Longxiang Tao, Fengqiong Yu, Yanghua Tian, et al., *Neural correlates of auditory verbal hallucinations in schizophrenia and the therapeutic


[64] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez, Mining large-scale smartphone data for personality studies, Personal and Ubiquitous Computing 17 (2013), no. 3, 433–450.


[80] Alex W DaSilva, Jeremy F Huckins, Rui Wang, Weichen Wang, Dylan D Wagner, and Andrew T Campbell, *Correlates of stress in the college environment uncovered by the application of penalized generalized estimating equations to mobile sensing data*, JMIR mHealth and uHealth 7 (2019), no. 3, e12084.


260


incentives based on three smartphone sensing studies, Social Psychological and Personality Science 8 (2017), no. 5, 479–492.


BIBLIOGRAPHY

[157] RE Hoffman, M Varanko, J Gilmore, and AL Mishara, *Experiential features used by patients with schizophrenia to differentiate ‘voices’ from ordinary verbal thought*, Psychological medicine **38** (2008), no. 8, 1167–1176.


[163] Jeremy F Huckins, Alex W DaSilva, Rui Wang, Weichen Wang, Elin L Hedlund, Eilis I Murphy, Richard B Lopez, Courtney Rogers, Paul E Holtzheimer,


BIBLIOGRAPHY


BIBLIOGRAPHY


[308] Runar Elle Smelror, Josef Johann Bless, Kenneth Hugdahl, and Ingrid Agartz, *Feasibility and acceptability of using a mobile phone app for characterizing audi-


[329] Vincent W-S Tseng, Akane Sano, Dror Ben-Zeev, Rachel Brian, Andrew T Campbell, Marta Hauser, John M Kane, Emily A Scherer, Rui Wang, Weichen


[341] Rui Wang, Min S. H. Aung, Saeed Abdullah, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill, Emily A. Scherer, Vincent W. S. Tseng, and Dror Ben-Zeev, *Crosscheck: Toward pas-


[350] Weichen Wang, Subigya Nepal, Jeremy F. Huckins, Lessley Hernandez, Vlado Vojdanovski, Dante Mack, Jane Plomp, Arvind Pillai, Mikio Obuchi, Alex daSilva, Eilis Murphy, Elin Hedlund, Courtney Rogers, Meghan Meyer, and


