Inference of Anxiety and Depression from Smartphone-Collected Data

by

Daniel Di Matteo

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Graduate Department of Electrical and Computer Engineering
University of Toronto

© Copyright 2021 by Daniel Di Matteo

Inference of Anxiety and Depression from Smartphone-Collected Data

Daniel Di Matteo
Doctor of Philosophy
Graduate Department of Electrical and Computer Engineering
University of Toronto
2021

Abstract

Anxiety and depression are widespread and have significant impact on individuals' quality of life. Automated technology which objectively measures behaviors relevant to symptoms of anxiety and depression could aid in the diagnosis and treatment of these disorders in two ways. Diagnosis could be improved by using this data to build automated pre-screeners for these disorders. Treatment could be improved by sharing relevant patient data with clinicians, aiding in decision making. Smartphones offer a wide variety of objective digital data, such as GPS location and audio, which can be collected and analyzed in order to infer key behaviors which give insight into individuals' mental health. This thesis describes the development of passive and objective measures of individuals' symptoms of anxiety and depression, enabled by the analysis of smartphone-collected data. A study was conducted where data was collected from participants' smartphones and a number of features which quantify symptom severity of anxiety and depression were designed and extracted from this data. One such feature, which infers the regularity in an individual's daily pattern of activity from short recordings of their ambient audio, was found to be correlated with self-reported measures of depression ($r = -0.37$, $P = 0.01$). Another feature, which infers the number of times that an individual leaves the home from GPS location data, was found to be correlated with self-reported measures of social anxiety ($r = -0.25$, $P = 0.04$), generalized anxiety ($r = -0.31$, $P = 0.01$), and depression ($r = -0.29$, $P = 0.01$). In total, 97 features were extracted from participant data and tested for associations with self-report measures of anxiety and depression. A subset of these features were also selected and used to build predictive models of anxiety and depression, screening participants for social anxiety disorder, generalized anxiety disorder, and depression with mean AUROC of 0.67, 0.55, and 0.73, respectively. The results of this study show that objective data, passively collected from individuals' smartphones, give broad insight into individuals' behaviors and activities, and specific insight into the severity of symptoms of anxiety and depression.

To my parents

Acknowledgments

This work was supervised by Professor Jonathan Rose, of the University of Toronto, and Dr. Martin Katzman, Psychiatrist and Director of the START Clinic. Professor Rose, thank you for giving me the freedom and support to take ownership of this work — doing so, under your patient guidance, I have learned to become a professional and an engineer. Doctor Katzman, thank you for your mentorship and for instilling in me the confidence that enabled me to undertake this work. Working in this unfamiliar territory would not have been possible without all your help.

Thank you to everyone at the START Clinic who helped me with this research: Tia Sternat, Kathryn Fotinos, Alexa Fine, Sachin Lokuge, Emilia Sherifi, Julia Yu, Geneva Mason, and Teresa Chaves. Thank you Wendy Wang for your interest and work on this project. Thank you to all the other grad students in Professor Rose's research group who gave me feedback and guidance throughout my journey, especially Henry Wong, Jason Luu, Alex Rodionov, and Braiden Brousseau, who all showed the new guy how to do things properly.

A special thanks goes out to Jonathan Polak: thank you for introducing Professor Rose and I to this revolutionary way of bridging mental health and technology. Your vision put us on the path that resulted in this work.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The past decade has seen an increase in advocacy and action surrounding mental health problems in Canada and abroad. The reasons for this are clear: the problems caused by mental illness are widespread and severe. It is estimated that one in five Canadians suffer from some form of mental illness annually, and that by the year 2041 more than eight million Canadians will be living with a mental illness [1]. The impact of mental illness are felt by both individuals and society. Individuals who suffer from a mental illness have a lower quality of life [2] and a shorter life expectancy [3]. The economic impact of mental illness to the Canadian economy is estimated to be $51 billion per year [4].

While the challenges posed by mental illness are considerable, there is hope for those who suffer. Mood and anxiety disorders, which are the most common classes of mental illnesses in Canada, can be successfully treated with psychotherapy, pharmacotherapy, or a combination of both [5]. The success of both of these treatment modalities hinges upon individuals having access to mental health practitioners and on practitioners having key information about their patients' conditions.

Meeting these two key requirements in mental health care practice —gaining access to practitioners and transmitting relevant information to them —are challenges in and of themselves. Access to diagnosis and treatment is lacking: it is common for many Ontario residents with mental health problems to wait for 6 months to 1 year for treatment [6]. Once treatment begins, however, the difficulty involved in patient-clinician communication also begins. A clinician only learns symptoms and problems typically from the patient, who is often unable or unwilling to provide accurate detail on their condition and behaviour. The short and infrequent nature of consultations between patients and clinicians (usually 1 hour, biweekly) makes the problem of diagnosing and assessing a patient's progress in treatment even more difficult.

Some form of technology which could track subjects' symptoms of mental health

problems in an objective fashion, over long periods of time, could potentially address both of these issues, in different ways. It might help shorten wait times for diagnosis and expedite treatment, as it could be deployed as a screener for the diagnosis of mental health disorders . It could also improve clinicians' abilities to diagnose, treat, and detect relapses in mental illness by providing them with useful supplementary data about their patients' history of symptoms over time.

The simplest form of technology for screening and long-term symptom tracking of mental illness is simply pen-and-paper. A number of questionnaires exist which can be employed as mental health screeners [7]–[9], and journaling-based approaches to symptom tracking also exist [10]. We see three problems with this approach. Firstly, it is not easily a *persistent* method of tracking mental state, since individuals can simply forget or lose the motivation to log the necessary information. Secondly, it is not *objective*, since it requires individuals to self-assess their subjective feelings and state of mind. Third, it is not *passive*, in that it requires individuals to actively engage in the act of writing, something that is not easily done on-the-fly in different day-to-day environments.

This work seeks to make use of the sensors and information found in modern smartphones in order to infer and track a subject's symptoms of mental illness. The widespread use of smartphones, coupled with their persistent network connectivity, makes them a convenient alternative to any form of dedicated sensors or device that would otherwise be obtrusive and unnatural for the patient. The common sensors present on all modern smartphones offer a rich landscape of data one can use to infer behaviour or actions pertinent to one's state of mind. Identifying long periods of inactivity on the phone (or lack thereof) can be used as an estimate of the patient's sleep patterns. Tracking GPS location offers clinicians insight into how often a patient is leaving the home. Having access to the ambient audio throughout a patient's day, specifically conversational speech, is of great interest. Consider, for example, a patient being treated for social anxiety. The number of spoken conversations that the patient has participated in is relevant to clinicians when determining whether treatment is effective or not.

In contrast with simple pen-and-paper tracking, smartphone-based tracking has three attractive properties for use in the inference of mental health: it can be collected in a manner that is *persistent*, *objective*, and *passive*. Persistent data collection is important because it offers insight into subject's state across all points in time. Given that most people now take their smartphones with them everywhere they go, smartphones are an ideal platform for persistent data capture. The sensors which collect the data are also objective instruments, as opposed to subjects' memories

and feelings. Finally, smartphone data can be collected in a manner that is entirely passive, with no interaction from the user. This helps to minimize the potential for the measurement tool itself to impact the mental state of the subject being assessed. Active measurement, such as asking direct questions, could potentially influence a person's state of mind or induce a placebo effect, as it causes subjects to consistently engage with and reflect upon their disorder.

1.1 Focus and Goals

The broad goal of this work is to infer individuals' mental health in an automated fashion from smartphone-collected data. There are numerous classes of mental health disorders, and many aspects of how these disorders present and impact people's lives could be studied. This work focuses upon one of the most common mood disorders, depression, and two common anxiety disorders, social anxiety disorder and generalized anxiety disorder. Given this focus, our goal is to determine if it is possible to screen for social anxiety disorder, generalized anxiety disorder, and depression, in a group of study participants, using only data collected from their smartphones.

1.2 Contributions

In order to achieve the goal stated above, this work makes the following contributions:

- A study was conducted to explore the willingness of patients to consent to smartphone-based assessment of mental health
- The creation of a complete data collection system consisting of an Android smartphone application, backend infrastructure, and study enrollment website
- A general methodology which allows for data collection from human participants in a manner that is remote, anonymous, and automated
- The generation of novel features of anxiety and depression, including a number of features derived from recordings of ambient environmental audio
- The prediction of generalized anxiety disorder, social anxiety disorder, and depression, with mean AUROC of 0.67, 0.55, and 0.73, respectively

1.3 Organization

The remainder of the thesis is organized as follows: Chapter 2 provides relevant background information on the classes of mental disorders investigated in this work, and

also provides a review of the existing studies which have attempted to infer mental health from digital data. Chapter 3 presents our assessment of individuals' willingness to allow a smartphone application to collect their personal data and assess their mental health. Chapter 4 details the study which was run to deploy the smartphone data collection system to participants. Chapter 5 gives a description of the smartphone-based data collection system which was built and used in this work. Chapter 6 gives a high-level summary of the demographic and smartphone data collected from study participants. Chapters 7 through 9 describe how predictors of mental health, referred to as *features*, were engineered and extracted from participants' smartphone-collected data. Chapter 10 presents the results of using these features to predict participants' mental health. Finally, the thesis is concluded in Chapter 11, where the work is summarized and a roadmap for future work is provided.

Chapter 2

Background and Related Work

This chapter is organized into two sections. The first section provides some necessary background information on the classes of mental disorders which are under investigation in this work. The second section gives a survey of relevant work in the literature which has sought to infer the presence of these same classes of mental disorders from individuals' digital data.

2.1 Anxiety and Depression

Anxiety disorders and depression, as psychiatric disorders, are defined and diagnosed in terms of clusters of symptoms [11]. These symptoms are phenomena that arise from an underlying complex of cognitive, emotional, behavioral, and physiological processes, some of which are not objectively measurable or observable (i.e., in the case of emotional processes) [11]. This is in contrast with a more idealized system, where classification and diagnosis is performed on the basis of some objectively measurable characteristics of the body, which can include so called biological markers or "biomarkers". An example of such a biomarker would be LDL cholesterol levels in the blood as a biomarker of risk for cardiovascular disease [12]. Research is underway to identify biomarkers for mental disorders, but their ability to be used for the diagnosis of anxiety disorders or depression has yet to be demonstrated [13], [14]. As such, one of the key motivations behind this research is to determine whether relevant behaviours, activities, and states of mind can be inferred in an automated fashion from smartphone-collected data, producing what are sometimes referred to as digital biomarkers [15].

The specific set of behaviours and symptoms that are used to define and diagnose a mental disorder are referred to as diagnostic criteria. Diagnostic criteria are specified by two widely established classification systems: the International Classification of

Diseases (ICD) [16], produced by the World Health Organization and used worldwide, and the Diagnostic and Statistical Manual of Mental Disorders (DSM) [17], produced by the American Psychiatric Association. The DSM is the most popularly used system in the US and Canada and will be used as the reference in this work. This section provides some background on the key characteristics and diagnostic criteria of the three disorders under focus in this work: social anxiety disorder, generalized anxiety disorder, and major depressive disorder. We also identify the instrument which will be used to screen for and measure the severity of each disorder. Finally, this section introduces a general measure of mental health that was considered in parallel with the three aforementioned disorders.

2.1.1 Social Anxiety Disorder

Social Anxiety Disorder (SAD), sometimes referred to as Social Phobia, is a mental disorder characterized by feelings of fear, nervousness, or anxiety surrounding situations in which one may be judged, evaluated, or scrutinized by others [11]. These situations include, but are not limited to, eating in public, meeting new people, public speaking, being the centre of attention, or attending parties [11]. While it is a commonly shared experience to feel nervousness in social situations or to fear judgment at times throughout one's life, social anxiety disorder is distinguished from more benign shyness by the degree to which individuals avoid social situations. Social interactions can enrich one's life, and therefore severe avoidance of social situations causes disability in individuals suffering from the diagnosis. Those who suffer from SAD may also be afraid of showing signs of fear or anxiety that may be perceived by others, for example, by being seen trembling, blushing, or sweating. It is estimated that 8% of Canadians will suffer from SAD at some point in their life [18], although other measures have suggested the lifetime prevalence may be as high as 13% [19].

Measuring the Severity of SAD

The assessment strategies used to diagnose SAD are often built from a series of dichotomous (present/not present) questions, where each question directly attempts to discover whether a specific diagnostic criterion is present. As such, it can be argued that they do not provide a clear mechanism to measure exactly how severe a particular instance of SAD is. A problem with this is that it is understood that the individual symptoms underlying the disorder can vary in severity. A dimensional rating (i.e., a rating scale measured at interval level) therefore captures more information. A severity measure enables clinicians and researchers to capture individual differences in disorder severity, either between patients, between patient populations, or across

time for a single patient or group of patients [20]. Furthermore, severity measures can be used as treatment outcome variables for studies, for example, in studies where researchers attempt to measure the effect of treatments for the disorder [20].

As with diagnosis, severity measures for SAD are also descriptive, and for the same reasons. These severity measures, often referred to as rating scales, rate how a patient feels or behaves in a given set of circumstances, and then combines these individual ratings into a final score or severity measure. Clinicians will generally ask the patient, for example, how anxious they may feel when speaking in public, as rated on a 4-point ordered scale with options such as "not at all", "somewhat", "moderately", and "extremely". These scales are designed to cover a wide enough set of scenarios, feelings, and behaviours so as to get an assessment of how severe a patient's particular instance of the disorder is in a general sense.

Many rating scales are designed to be rated by an interviewing clinician. In a manner similar to diagnosis, clinicians are free to assess the individual questions on the scale in an unscripted manner in order to best interact with a particular patient. There also exist self-report forms of many rating scales in which the patient independently reads and answers the questions comprising the scale, in questionnaire form. Many self-report scales have been demonstrated to have reliability and validity comparable to clinician-administered forms when completed by the patient. Some clinically validated self-report questionnaires to assess SAD severity are the Liebowitz Social Anxiety Scale (LSAS) [21], the Social Phobia and Anxiety Inventory (SPAI) [22], the Social Phobia Inventory (SPIN) [23], the Social Phobia Scale (SPS) [24], and the Social Interaction Anxiety Scale (SIAS) [24]. The social anxiety instrument used in this work is the Liebowitz Social Anxiety Scale.

The Liebowitz Social Anxiety Scale

The Liebowitz Social Anxiety Scale (LSAS) is the most commonly used scale to measure the severity of SAD [20], [25]. It consists of 24 multiple-choice questions, where each question presents the subject with a hypothetical situation in which they may find themselves and asks the subject to rate how often they avoid the situation and how much they fear the situation (both ratings on a 4-point ordinal scale). The situations in the questions are either representative of social activities (13 of 24 questions) or of situations in which the subject's performance will potentially be evaluated by others (11 of 24 questions). In addition to the total score, the LSAS can be subdivided into 6 sub-scores: fear of social interaction, avoidance of social interaction, fear of performance situations, and avoidance of performance situations, total fear, and total avoidance. A sample of the LSAS can be viewed at [26].

The LSAS can be administered by clinicians or completed in self-report form. The self-report version of the LSAS has been demonstrated as having validity and reliability similar to that of the clinician-administered form [27]. Paper and pencil completion of the self-report LSAS is by far the most common method, however the study in [28] has shown that the scale can also be administered electronically by PC, again yielding reliability, consistency, and validity comparable to the traditionally-administered form. The LSAS also functions well as a screener for social anxiety disorder, where a threshold score of 60 or greater screens for the generalized subtype of social anxiety disorder with 82% sensitivity and 78% specificity [29]. Sensitivity and Specificity are metrics used to assess the accuracy of a diagnostic tool or screener, and are defined as follows:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2.1}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \tag{2.2}$$

A True Positive refers to an individual who has the disorder in question and was correctly identified as having the disorder by the screener. Similarly, a True Negative refers to an individual who does not have the disorder in question and was correctly identified as *not* having the disorder by the screener. False Positives and False Negatives are both misclassifications. A False Positive refers to an individual who was identified as having the disorder when they, in fact, do *not* have the disorder. A False Negative refers to an individual who was identified as not having the disorder when they, in fact, *do* have the disorder. Therefore, a screener with high sensitivity has a high probability of detecting a disease or disorder if present, while a screener with high specificity has a high probability of correctly detecting the *absence* of a disease or disorder.

2.1.2 Generalized Anxiety Disorder

Generalized Anxiety Disorder (GAD) is a psychiatric disorder characterized by excessive worry. The worry is difficult to control, and is accompanied by a number of additional symptoms including restlessness, fatigue, difficulty concentrating, irritability, muscle tension, and disturbed sleep [17]. The worry which characterizes generalized anxiety disorder is often associated with a number of everyday life scenarios, such as finances, job responsibilities, or worrying about friends and family [17]. A diagnosis of generalized anxiety disorder requires that the worries are not better explained by some other disorder. For example, social fears would correspond more closely with

social anxiety disorder, or a single specific object or situation with a specific phobia. As is the case with social anxiety disorder, the worry or fear felt is out of proportion with the threat presented by the object of worry and causes significant distress to the individual. The distress associated with their anxiety disorder symptoms negatively impacts the individual's quality of life, which can affect their social, professional, or family lives, and often prevents them from performing specific activities [17]. It is estimated that 8.7% of Canadians will suffer from generalized anxiety disorder at some point in their life [30].

A number of rating scales can be used, either in clinician-administered or self-report form, to assess the severity of generalized anxiety disorder. One clinician-administered rating scale is the Generalized Anxiety Disorder Severity Scale (GADSS) [31]. Self-report measures include the Hospital Anxiety and Depression Scale - Anxiety (HADS-A) [32], and the GAD-7 7-item anxiety scale (GAD-7) [8]. The generalized anxiety instrument used in this work is the GAD-7.

GAD-7

The GAD-7 is a commonly used self-report measure of generalized anxiety disorder. It consists of 7 questions, each graded on a 4-point Likert scale, resulting in a total score which can range from 0 to 21. The questions ask respondents to rate the severity of their feelings of anxiety and worry over the past two weeks. Each question in the GAD-7 measures the severity of one of the symptoms of generalized anxiety disorder. The instrument has been shown to have good reliability and validity [8]. The GAD-7 also functions well as a screener for generalized anxiety disorder, where a threshold score of 10 or greater achieves 89% sensitivity and 82% specificity [8]. The GAD-7 is a good alternative to many of the other anxiety measures as it is faster to administer, due to its relatively small number of items. A copy of the GAD-7 is included in Appendix A.

2.1.3 Major Depressive Disorder

Major Depressive Disorder (MDD) is characterized by the presence of what are technically referred to as major depressive episodes. These major depressive episodes find individuals in a state where they experience low mood, diminished interest or pleasure in activities, changes in weight, disturbances to sleep, fatigue and low energy, diminished ability to concentrate, and may have thoughts of death and/or desire to commit suicide [17]. The persistence of these symptoms over at least a two-week period qualifies for a major depressive episode. The presence of a major depressive episode with no disturbances to mood that are associated with other mood disorders

(e.g., bipolar disorder) then qualifies as major depressive disorder. It is estimated that 9.9% of Canadians will suffer from major depressive disorder at some point in their life [33].

A number of rating scales can be used, either in clinician-administered or self-report form, to assess major depressive disorder. Some scales include the Hamilton Depression Rating Scale (HAM-D) [34], Beck Depression Inventory (BDI) [35], the Montgomery-Asberg Depression Rating Scale (MADRS) [36], the Patient Health Questionnaire-9 (PHQ-9) [9], and an eight-item version of the PHQ-9 known as the PHQ-8 [37]. The depression instrument used in this work is the PHQ-8.

PHQ-8

The PHQ-8 instrument is a brief, self-report rating scale for assessing depression. Each item is scored on a 4-point Likert scale, resulting in a total score which can range from 0 to 24. The PHQ-8 is identical to the PHQ-9 instrument, with the exclusion of the last item, which is a question which assesses whether the respondent has considered self-harm or suicide. The PHQ-8 was selected for this work, instead of the PHQ-9, because the remote and anonymous nature of our study would not allow us to properly intervene in the case where participants demonstrated suicidal ideation (as assessed by the last iten of the PHQ-9).

The nine items of the PHQ-9 each correspond to one of the 9 diagnostic criteria which define a major depressive episode, each asking the respondent how often they have experienced each criterion over the past two weeks, from "not at all" to "nearly every day". The PHQ-9 and PHQ-8 can both be used to assess symptom severity and as a diagnostic screener for major depressive disorder. A PHQ-9 score at or above 10 achieves 88% sensitivity and 88% specificity when screen for major depressive disorder [9]. The same threshold score applied to PHQ-8 scores was also shown to be an effective screening criterion [37]. A copy of the PHQ-8 is included in Appendix A.

2.1.4 General Functional Impairment

Many mental disorders exert a marked impact upon quality of life, in part due to functional impairment. Unlike social anxiety disorder, generalized anxiety disorder, and major depressive disorder, functional impairment is not a class of mental illness and there is no associated diagnostic scheme. It is a useful construct, however, in this study, since all three disorders under study do have the potential to impair the ability of individuals to perform everyday tasks. As such, having a measure of general functional impairment helps to lend further validity to the study methods, since any inferred mental illness should be accompanied by some degree of functional

impairment. To put it another way, mental disorders are associated with lower quality of life, and another way to measure impacts to quality of life are through measures of functional impairment. A commonly-used measure of impairment due to mental illness is the Sheehan Disability Scale (SDS) [38].

Sheehan Disablity Scale

The Sheehan Disability Scale is a 5-item, self-report measure of impairment. The first three items ask respondents to rate to what degree their symptoms have impaired their work, social, and home life, respectively, over the past week. These three items are assessed on a 10-point visual-analog scale. The final two items assess how many days, over the past week, they have lost and been unproductive as a result of their symptoms. The SDS is commonly scored by simply summing the first three items, resulting in a score which ranges from 0 (unimpaired) to 30 (highly impaired). A sample of the SDS can be viewed at [39].

2.2 Inference of Anxiety and Depression from Digital Data

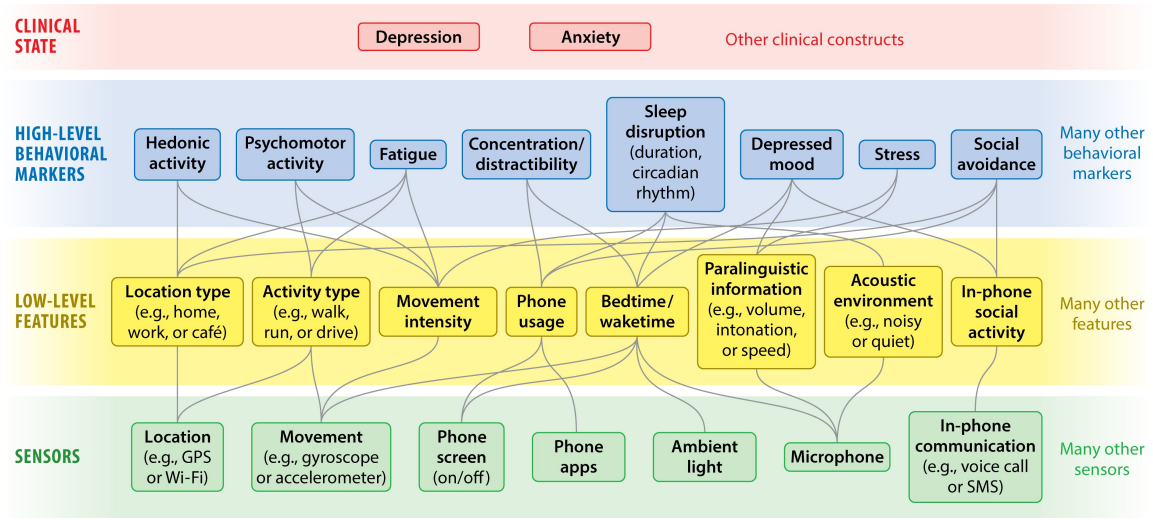
A core aspect of this work is the extraction of knowledge relevant to mental health from the wide array of smartphone-collected data. To convey a sense of scale of this class of data, consider that smartphones can sense, measure and record geolocation, motion, ambient lighting conditions, ambient audio, communication patterns and app usage, all in real-time. In a matter of weeks, tens of thousands of data points across these many data modalities can easily be collected from a single individual. It is challenging to build predictive models that are both accurate and also offer insight into what specific patterns of data, and what points in time, are driving models to make specific decisions or characterizations [40]. It is our goal with this work is to do just that.

A general approach to this problem of knowledge extraction and interpretability has been to review the diagnostic criteria and the key symptoms of a mental health disorder and look for data streams that would relate to these. Thus the question is "can the presence or absence of a symptom or characteristic of a disorder be inferred from the data?", and in this way can we transform data into relevant knowledge. For example, key symptoms and criteria for depression include sleep disturbances (insomnia or hypersomnia) [41] and low energy [42]. The duration of sleep can be inferred by analyzing a combination of light and motion sensor data, screen activity, and ambient audio [43]. While a person's energy level is difficult to measure directly, proxy measures can be created by observing the degree to which subjects leave the

home and travel throughout their environments [44], or by observing activity levels as measured by motion sensors [45].

This approach to knowledge extraction for the inference of mental health has been conceptualized as a hierarchical framework by Mohr et al. in [46], and is illustrated in Figure 2.1.

Figure 2.1: Knowledge extraction framework for inference of mental health [46]



AR Mohr DC, et al. 2017.
Annu. Rev. Clin. Psychol. 13:23–47

The lowest level of the hierarchy is composed of the wide array of sensors found in smartphones and the associated data that they produce. These data can then be used to compute low-level features, which capture aspects of behavior pertinent to different aspects of mental health. The value in features is that they can be used, either directly or in combination with other features, to construct higher-level *behavioral markers*, which are objective measurements of the types of behaviors and states which are known (or hypothesized) to be associated with clinical state. For example, a "Movement Intensity" feature can be computed from "Movement" sensors (i.e., motion sensors). Given an individual's Movement intensity, one could then infer their latent psychomotor activity and fatigue, two key criteria for evaluating depression, which is one of clinical states that we ultimately wish to measure. These high-level behavioral markers are referred to as latent, because, in general, these are conceptual constructs which cannot be measured directly, but can be inferred from the lower-level features. With the exception perhaps of Sleep disruption and Psychomotor activity, these behavioral markers do not have direct observable states that can be measured objectively. While concepts like fatigue and mood are clearly relevant and

critical to defining and measuring clinical state, they do not lend themselves to direct quantification in an objective form, as opposed to the low-level features, which do.

It is worth noting that, in practice, the high-level behavioral markers are not often computed, but are presented as concepts which serve to justify the design of low-level features, which are then used to directly infer clinical state. For example, continuing to refer to Figure 2.1, a researcher may propose to use a combination of the "Location type", "Phone usage" and "In-phone social activity", features to infer social anxiety, since all these features could give insight into the degree of "Social avoidance" (a high-level behavioral marker) that an individual may be displaying. Given that more in-phone social activity might signal less social avoidance, this is a potentially relevant feature for the inference of social anxiety disorder, which is characterized by social avoidance. However, at no point in the actual mechanics of the inference is a "Social avoidance" marker being directly and intentionally computed. The inference of clinical state (social anxiety, in this example) would simply proceed using low-level features directly as input, and if the inference succeeds to build a model which has predictive capacity of clinical state, it does so on the basis of the "link" from feature to latent marker and latent marker to clinical state.

The prior work on inferring mental health from smartphone-collected data can all be viewed through this framework, where features are extracted from digital data and mental state is inferred from these features by way of the association from features to clinical state (acting "through" latent behavioral markers). The following subsections provides a overview of this prior work. While our own study focuses directly upon the inference of anxiety (social anxiety and generalized anxiety) and depression, there is some relevant work that does not focus on these strict clinical disorders but is nonetheless relevant and useful to review. As such, studies which have investigated the prediction or inference of stress and worry will also be considered along with those that focus specifically upon anxiety disorders, as stress and worry are characteristic features of this class of disorders. Similarly, for the investigation of works centered on depression, we will include some studies which did not specifically focus on clinical depression, but did investigate low or depressed mood. For all works included here, the focus is upon methodologies which used smartphone-collected data, or data which could be collected using smartphones or mobile devices.

2.2.1 Studies of Anxiety, Stress, and Worry

Wang et al., in the StudentLife study [47], investigated associations between smartphone-collected data and self-reported symptoms of depression, stress, flourishing, and loneliness. Participants included 48 college students, who were enrolled in a 10 week study.

Features which were found to be associated with stress were conversation frequency (r = -0.394, $P = 0.013$), conversation duration (r = -0.357, $P = 0.026$), and sleep duration (r = -0.355, $P = 0.024$). These results show that more social interaction and sleep were all both associated with lower levels of stress in the study participants. Conversation frequency and duration were both measured in a novel way, by detecting the presence of speaking voices in recordings of ambient audio recorded by participants' smartphones. Sleep duration was inferred from a combination of smartphone-collected accelerometer data, sound recordings, light levels, and screen activity.

Ben-Zeev et al. [48] have also investigated the association between smartphone-collected data and subjective stress. In a study of 47 individuals, the authors found that total distance travelled and sleep duration were both inversely associated ($P < .05$) with subjects' self-reported levels of stress over the 10-week study period. Daily deviations in distance travelled (relative to the mean distance travelled over the entire 10 weeks) was also show to be negatively associated ($P < .05$) with daily stress. Distance travelled was computed from a combination of smartphone-collected GPS data and WiFi access point connections over time. Sleep duration was inferred from a combination of smartphone-collected accelerometer data, sound recordings, light levels, and screen activity. Subjects' stress levels were quantified using the Perceived Stress Scale [49], a self-report stress assessment instrument.

Chow et al. [50] have investigated the relationship between affect and time spent in the home. Motivated by the known tendency of individuals experiencing negative emotions to withdraw socially, they hypothesized that time spent at home would be associated with more negative affect, which was partly quantified by a self-reported measure of social anxiety (the Social Interaction Anxiety Scale [51]). In their study consisting of 72 undergraduate subjects, time spent at home was measured from smartphone-collected GPS data. The results confirmed their hypothesis, and it was shown that higher self-reported levels of social anxiety were associated with more time spent at home (standardized beta=.05, P=.007).

Boukhechba et al. [52] have investigated the association between social anxiety levels and a number of GPS-data-derived features. Study participants included 228 undergraduate students, who supplied self-reported social anxiety levels and smartphone-collected data over a 2-week trial. A key contribution of this study is the use of so-called "semantic location" data, where specific geolocations are labelled in terms of their type. Some types include, for example, leisure (cinemas, pubs, etc), education (schools, libraries), and home (the location at which an individual spends the most time). This semantic location is of great interest in studies of social anxiety

since social anxiety is characterized by the avoidance of situations (and therefore locations) where social interaction is expected to occur. Key findings were that highly socially anxious individuals visited food places less frequently and also spent more time at home after school hours. Using a combination of 4 classes of location-based features, they were also able to classify participants as high or low socially anxious with an accuracy of 86%. These features were (1) a measure how much time was spent at different location types, (2) the distribution of visits, over time, to each location type , (3) a measure of how participants' chose to spend their time between different location types, and (4) a feature which characterized how often a participant transition from each possible pair of location types.

Summary

- Greater *conversation frequency* and *conversation duration* is associated with less stress [47]
- Greater *sleep duration* is associated with less stress [47], [48]
- Greater *distance travelled* is associated with less stress [48]
- More *time spent at home* is associated with stronger symptoms of social anxiety [50], [52]
- Fewer *visits to social or busy locations* is associated with stronger symptoms of social anxiety [52]

2.2.2 Studies of Depression and Low Mood

The study by Wang et al. [47] referred to in Section 2.2.1, in addition to investigating symptoms of stress, also investigated self-reported levels of depression as measured by the PHQ-9 instrument. Features which were found to be associated with study participants' PHQ-9 scores were sleep duration (r = -0.382, $P = 0.020$), conversation frequency (r = -0.403, $P = 0.010$), conversation duration (r = -0.328, $P = 0.044$), and number of encounters with other people (r = -0.362, $P = 0.025$). Conversation frequency and duration were both measured in a novel way, by detecting the presence of speaking voices in recordings of ambient audio recorded by participants' smartphones. Sleep duration was inferred from a combination of smartphone-collected accelerometer data, sound recordings, light levels, and screen activity. The number of encounters that participants had with other individuals was inferred through the number of times that their smartphone encountered another Bluetooth-enabled device. These results

indicate that, similar to stress, more social interaction and more sleep are associated with lower levels of self-reported depression.

A second study by Wang et al. [53] further investigated depression using smartphone-collected data. In this 7-week study of 84 students, a number of features computed from smartphone collected data were found to be correlated with self-reported depression (measured by the PHQ-8 instrument). As a proxy measure of the inability to focus which characterizes depression, they show that students who used their phone more often during the day had higher PHQ-8 scores (r = 0.282, $P = 0.010$). They show that stronger depressive symptoms are associated with greater variation in bed time (r = 0.301, $P = 0.024$) and wake time (r = 0.271, $P = 0.043$). Furthermore, more time spent stationary was associated with stronger depressive symptoms (r = 0.374, $P = 0.009$), while a greater number of places visited was associated with weaker depressive symptoms (r = -0.269, $P = 0.023$). The authors also built predictive models which were able to detect depression, defined as having a PHQ-8 score above 10, with AUC = 0.809.

The study by Ben-Zeev et al. [48] referred to in Section 2.2.1, in addition to investigating symptoms of stress, also investigated self-reported levels of depression as measured by the PHQ-9 instrument. Two features which were found to be significantly associated with stress were also found to be associated with depression: total distance travelled and sleep duration. Increased distanced travelled was associated with lower symptoms of depression. The association between sleep duration and depressive symptoms differed over time, with more sleep time first being associated with stronger symptoms of depression in the first 45 days of the study, but then with decreased symptoms after day 45. Speech duration was also found to have a significant association with self-reported depression. The association between speech duration and depression also varied across time, with more speech being associated with lower depressive symptoms initially in the study, but later being associated with greater symptoms of depression in the later days of the study.

Canzian and Musolesi [54] investigated the link between self-reported depression and smartphone-collected GPS location data. Study participants included 28 individuals who were monitored for, on average, 71 days, and were asked to provide daily self-reported measures of depression via the PHQ-8 instrument. The authors designed and extracted eight location-based features. These location-based features were extracted from GPS location on a daily basis, and the correlation between the time series of daily features and daily PHQ-8 scores for each study participant were computed, and then averaged across all study participants. The feature they found to have the strongest average correlation with depression scores, maximum distance, is

the distance between the two most distant locations visited by an individual (this can be thought of as the "diameter" of the zone in which an individual travels over time). Maximum distance was found to have an average correlation with PHQ-8 scores of r = 0.432 ($P = 0.069$). Using all eight features in a predictive analysis, the authors were able to predict which days individuals reported elevated depressive symptoms with a sensitivity of 74% and specificity of 78%.

Saeb et al. [44] have also investigated the link between self-reported depression and smartphone-collected data. In a pilot study of 28 individuals, a number of location and phone-usage based features were found to be significantly correlated with self-reported depression severity (measured using the PHQ-9). The number of times that participants interacted with their phones was associated with stronger symptoms of depression (r = 0.52, $P = 0.015$). The amount of time individuals spent using their phone was also associated with stronger symptoms of depression (r = 0.54, $P = 0.011$). Some key location-derived features included *location variance*, a measure of variability in GPS location, *location entropy*, a measure of the variability of time spent at different locations, and *circadian movement*, a measure of the degree to which an individual's patterns of movement follow a 24-hour cycle. Higher location variance was associated with lesser symptoms of depression (r = -0.58, $P = 0.012$), as was location entropy (r = -0.58, $P = 0.012$) and circadian movement (r = -0.63, $P = 0.005$). These results indicate that more activity and more regularity in activity are all associated with weaker symptoms of depression. In a follow-up study, the strong association of these location-based features was replicated in a new sample of individuals [55]. Again, higher location variance was associated with lesser symptoms of depression (r = -0.43, 95% CI -0.437 to -0.423), as was location entropy (r = -0.44, 95% CI -0.445 to -0.435) and circadian movement (r = -0.48, 95% CI -0.486 to -0.474).

Many of the location-derived features proposed by Saeb in [44] were also used by Farhan et al in [56]. Similar results are reported, with correlation analysis indicating similar directional associations between the location-based features and self-reported depression symptoms (also using the PHQ-9). For example, location variance and location entropy were shown to be negatively correlated with PHQ-9 scores. The authors also conducted a predictive analysis to predict PHQ-9 scores from features. PHQ-9 scores of 25 Android-phone owning participants were predicted with sensitivity of 83% and specificity of 92%. PHQ-9 scores of 54 iPhone-owning participants were predicted with sensitivity of 73% and specificity of 97%.

A study by DeMasi et al. [45] sought to determine if a number of behavioral features derived from smartphone-sensed accelerometer data could predict changes in depressive symptoms. In their study of 44 participants, features which measured

sleep duration and the irregularity of time spent still were all found to be important predictors of self-reported symptoms of depression (measured by the Beck Depression Inventory). Greater variability in sleep duration was associated with greater depressive symptoms (linear regression coefficient $\beta = 7.2069$, $P < 0.001$). Greater variation in the amount of time in which individuals phones were still was associated with lesser depressive symptoms (linear regression coefficient $\beta =$ -3.3079, $P = 0.001$). Sleep duration was inferred from accelerometer data by noting the longest duration where the phone was still (i.e., no acceleration measured).

Pratap et al. [57] conducted a relatively large-scale study in predicting mood from smartphone-collected data of 271 participants. This study found generally weak relationships between features and mood as measured by a brief, 2-item self report measure of mood derived from the PHQ-9, called the PHQ-2. The strongest measured association among features and PHQ-2 scores was produced by a feature which measured distance travelled on foot or bicycle as measured by GPS (regression coefficient $\beta =$ –0.04, $P < 0.05$). Predictive modelling of the entire sample was generally unsuccessful, with median R^2 of the models being close to zero. The authors stress the need for personalized models by noting the large inter-subject variability of trends between features and PHQ-2 scores.

A final recent study by Chikersal et al. [58] has also attempted to detect and predict depression using smartphone-collected data. Study participants were 138 college students, enrolled in a 16 week study in which depressive symptoms were assessed by the Beck Depression Inventory-II (BDI-II). Features were extracted from GPS location, the presence of nearby Bluetooth devices, phone usage, call logs, step counting, and sleep tracking. Step counting and sleep tracking were performed by a paired wearable device (Fitbit). Numerous features were extracted from each stream of data, and each feature was extracted in 45 time slices throughout the study duration. Associations between features and depressive symptoms were not measured. A sophisticated feature selection technique was used in conjunction with randomized logistic regression to predict depressive symptoms (BDI-II score >= 14) with an accuracy of 85.7%.

Summary

- Greater *conversation frequency* and *conversation duration* is associated with weaker symptoms of depression [47]
- Greater *variation in sleep patterns* is associated with stronger symptoms of depression [45], [53]
- Fewer *places visited* is associated with stronger symptoms of depression [44], [53],

[55], [56]

- More *time spent at home* is associated with stronger symptoms of depression [44], [53], [55], [56]
- More *smartphone usage* is associated with stronger symptoms of depression [44], [53]
- More *physical activity* and less *time spent stationary* is associated with weaker symptoms of depression [53], [56], [57]

2.3 Summary

In Section 2.1 of this chapter we have given a brief summary of the key defining characteristics of social anxiety disorder, generalized anxiety disorder, and depression. Rating instruments for measuring the severity of these disorders and techniques for screening for these disorders were also discussed. In Section 2.2 of this chapter we have given a broad overview of the general methodology for the inference of mental health state from digital data, while also providing a review of prior work in the field along with a summary of key findings. The work described in this thesis builds upon prior work by utilizing smartphone-collected data sources that have yet to be analyzed, by creating a number of novel features from new and existing data sources, and also by applying this general methodology to the inference of generalized anxiety disorder.

In the following Chapter we describe the findings of an initial study where we sought to determine how likely individuals would be to provide the consent necessary to collect of their digital data in order to track and infer their mental health.

Chapter 3

Patient Willingness to Consent to Smartphone Data Collection

3.1 Introduction

The previous chapter has reviewed several approaches to inferring mental health state from data extracted from individuals' smartphones. While the reviewed inference techniques show promise in acting as novel, objective assessment tools, they rely upon the collection and processing of large amounts of private data from a patient's mobile phone. A critical requirement that is sometimes overlooked when discussing these techniques is that patients must first consent to this data collection. The decision to provide this consent is nontrivial, owing to the deeply personal and revealing nature of this data. Records of an individual's GPS location, for example, can reveal evidence of infidelity or substance abuse [59]. Furthermore, these records of location data and communications could result in health care providers being subpoenaed for these records by law enforcement agencies. As a result, patients may not be willing to consent to wholesale collection of these data for fear of potential social or legal ramifications. Patients may also feel uneasy knowing that their health care providers have the ability to scrutinize their actions and communications on a very fine-grained level. This is a specific concern for patients with anxiety disorder. In particular, for those with social anxiety disorder, where a source of anxiety is the fear of judgment from others, personal data collected from a mobile phone could potentially form the basis of that judgment. Furthermore, a patient may also fear the possibility that their data could accidentally become public and reveal confidential thoughts, feelings, and behaviors or that perhaps there could be legal implications if the government were to have access to the data [60].

For the researchers and practitioners interested in designing, experimenting with,

and deploying these kinds of mobile phone–based assessment tools, it will be important to have a sense of the patient population's willingness to consent to the necessary data collection. A number of studies have surveyed the general consumer population to determine the adoption of Internet of Things [61]–[63] and wearable technologies [64], [65] for health care purposes, and all clearly identified privacy concerns surrounding the data collected by these technologies. However, it is important to know specifically which sources of data available for collection on smartphones are of most concern and therefore least likely to achieve consent. This question was also raised by Torous et al. [66], who, in their study of patient interest in using mobile applications to monitor their mental health conditions, state that they did not address specifically to which sources of information patients would be willing to grant access. This information would allow researchers and developers to build systems that do not rely on unlikely-to-consent sources of data or to do extra work to find ways to address the concerns of patients on particular sources of data collection. It also gives insight into potential barriers that clinicians and researchers would need to address in order to deploy such systems in a health care setting.

The objectives of this part of the research were as follows: first, since smartphone–based assessment requires that patients own a smartphone and use it regularly, it was necessary to measure the ownership rates of smartphones within the patient population. We also sought to gauge the patient population's willingness to enroll in research studies in which their smartphone would be used as an experimental assessment tool for their mental health disorder. Finally, and most importantly, we sought to determine how likely patients would be to provide consent for each individual source of smartphone–collectible data across the wide variety of potential data sources.

3.2 Methods

3.2.1 Participants and Procedure

Participants included 82 individuals referred to the START Clinic for Mood and Anxiety Disorders, a tertiary care mood and anxiety disorder clinic, located in the city of Toronto, Ontario, Canada. Male and female genders were equally represented with 46% (38/82) males, 46% (38/82) females, and 7% (6/82) individuals who did not respond to a question on gender. The average age of the participants was 41 (SD 14.0) years old. All new intakes and existing patients (i.e., any patients in the waiting room) from August 2016 to October 2017 were recruited for the survey while waiting for their appointment with a clinician. Patients were asked to complete a

pen-and-paper questionnaire designed to achieve the goals stated above. The questionnaire underwent ethics review by Optimum Clinical Research (protocol number WS2382578), and all respondents provided informed consent before completing the questionnaire.

3.2.2 Materials

The questionnaire designed for this study consisted of 13 self-report questions. The full questionnaire as presented to respondents is shown in Appendix B. The introduction to the questionnaire provides context by presenting the concept of smartphone applications as a potential supplement or replacement to questionnaire-based methods of mental health assessment. It is explained that the collection and analysis of data from patients' mobile phones may assist their clinicians in providing better care yet may also impact their privacy. The questions, therefore, are to survey people's willingness to provide sources of information to clinicians and researchers in a scenario where a hypothetical application were installed onto their personal smartphone.

Questions 1 through 3 assess general smartphone use, ownership information, and willingness to install an application that might help with mental health. Questions 4 through 12 ask respondents if they would be willing to share a specific source of data available on their smartphone. The final question asks which specific brand of phone the respondent uses. For the two most potentially rich sources of information, SMS (short message service, or text) messages and raw ambient audio recordings made using the device's microphone, multiple questions are posed in which the amount of information and granularity of the data collection are varied. For example, question 5 asks if respondents would allow collection of SMS metadata (which doesn't contain the content of the message but surrounding information such as who the message was sent to/from and when message occurred), while question 6 asks respondents if they would allow analysis into the contents of their messages for the purposes of extracting potentially clinically relevant words.

Questions 10 through 12 are audio-related. Question 10, concerning the least detailed of the 3 audio data sources, asks respondents if they would share audio metadata, while question 11 concerns willingness to have speech recognition (word detection) performed upon their audio. Question 12 assesses willingness on the most detailed personal data: the unrestrained analysis of ambient audio.

3.2.3 Analysis

Responses to questions 1 through 12 were coded as ordinal variables with 2 (yes and no responses) or 3 levels (yes, maybe, and no responses). When respondents chose to use the other categories of response and provided free-form text, these responses were interpreted as either a yes, maybe, or no response and coded accordingly. Responses to question 13 (on the phone brand type) were coded as a categorical variable. To test for correlation between responses to questions and respondent age, the Spearman rank correlation coefficient (Spearman ρ) was computed along with P values to test against the alternative hypothesis that the correlation was nonzero (using the exact permutation test). To test for associations between responses to questions and respondent gender, responses were cross-tabulated by gender and a chi-square test for independence was performed (2-tailed). All statistics and tests were computed using MATLAB software version 2014b (MathWorks).

3.3 Results

Of the 82 respondents, 73 (89%) reporting owning a smartphone and using it daily (question 1). Rates of internet usage are also high, with 85% (70/82) of respondents reporting that they connect to the internet using their smartphone (question 2). All respondents reported owning either iPhone, Android, Blackberry, or Windows Mobile smartphones; Apple iPhones constituted 45% (35/78) of all smartphones, Android devices constituted 57% (37/78), Blackberry constituted 6% (5/78), and Windows Mobile the remaining 1% (1/78) (question 13). Regarding question 3, willingness to install and use a smartphone application to help their doctor better diagnose mental health problems and/or provide treatment, 41% of respondents (33/80) indicated that they would be willing to install and use such an application, with another 43% of respondents (34/80) indicating that they may be willing to use such an application, provided they were given more information about how it functioned. Only 16% respondents (13/80) indicated absolute unwillingness to use such an application. Table 3.1 presents how responses to these questions (questions 1, 2, 3, and 13) correlate with respondent age and how the responses are associated with respondent gender.

Survey questions 4 through 12 asked respondents if they would be willing to grant the hypothetical mental health monitoring application permission to collect a variety of data sources from their smartphone. The responses corresponding to each permission (data source) are presented in Table 3.2. Figure 3.1 provides a graphical representation of this data, along with respondent willingness to install the application. It is worth noting that very few of the responses to the survey questions were

Question topic	Correlation with respondent age		Association with respondent gender	
	ρ	P	χ^2	P
Mobile phone ownership (Q1)	0.08	0.50	2.81	0.24
Internet-connected mobile phone usage (Q2)	-0.09	0.46	1.58	0.21
Mobile phone types owned (Q13)	-0.10	0.43	1.52	0.68
Respondent willingness to install a mental health monitoring app (Q3)	0.10	0.44	1.86	0.39

Table 3.1: Correlation with respondent age and association with respondent gender for mobile phone statistics and general willingness to install a mental health monitoring application.

Permission	Willing to grant permission, n (%)			Correlation with respondent age		Association with respondent gender	
	Yes	Maybe	No	ρ	P	χ^2	P
GPS location (Q4)	28 (35%)	26 (33%)	26 (33%)	-0.07	.60	0.18	.91
SMS metadata (Q5)	24 (30%)	22 (28%)	34 (43%)	0.13	.31	1.00	.61
SMS contents (Q6)	16 (20%)	21 (27%)	42 (53%)	0.019	.13	6.30	.04
Calendar (Q7)	26 (33%)	26 (33%)	27 (34%)	-0.03	.83	4.73	.09
Screen on/off (Q8)	36 (46%)	18 (23%)	25 (32%)	0.00	.97	0.48	.79
Motion sensors (Q9)	33 (42%)	20 (26%)	25 (32%)	-0.04	.76	1.18	.55
Audio metadata (Q10)	16 (20%)	18 (23%)	46 (58%)	0.25	.04	4.58	.10
Audio keywords (Q11)	14 (18%)	20 (25%)	46 (58%)	0.29	.02	2.25	.33
Audio unrestrained (Q12)	15 (19%)	25 (31%)	40 (50%)	0.32	.01	4.67	.10

Table 3.2: Survey respondent willingness to grant permission by category.

strongly correlated with respondent age or associated with respondent gender. There is a moderate, positive correlation, however, between willingness to consent to audio recording and respondent age.

3.4 Discussion

3.4.1 Principal Findings

Smartphone ownership rates in the patient population surveyed in this study are high, with 89% of respondents reporting that they own a smartphone and use it daily and 85% of patients connect to the internet with it. These results indicate that, in general, lack of smartphone ownership itself will not present a barrier to the use of mental health monitoring applications in the future. This is further evidenced by American [67] and Canadian [68] mobile phone ownership statistics.

The vast majority of respondents reported owning either an iPhone (45%) or an Android device (47%). This corroborates surveys conducted by Gartner in 2018 [69] which indicate that Android-based and iOS-based smartphones, collectively, comprise 99.8% of the smartphone market. As each platform can require significant effort to

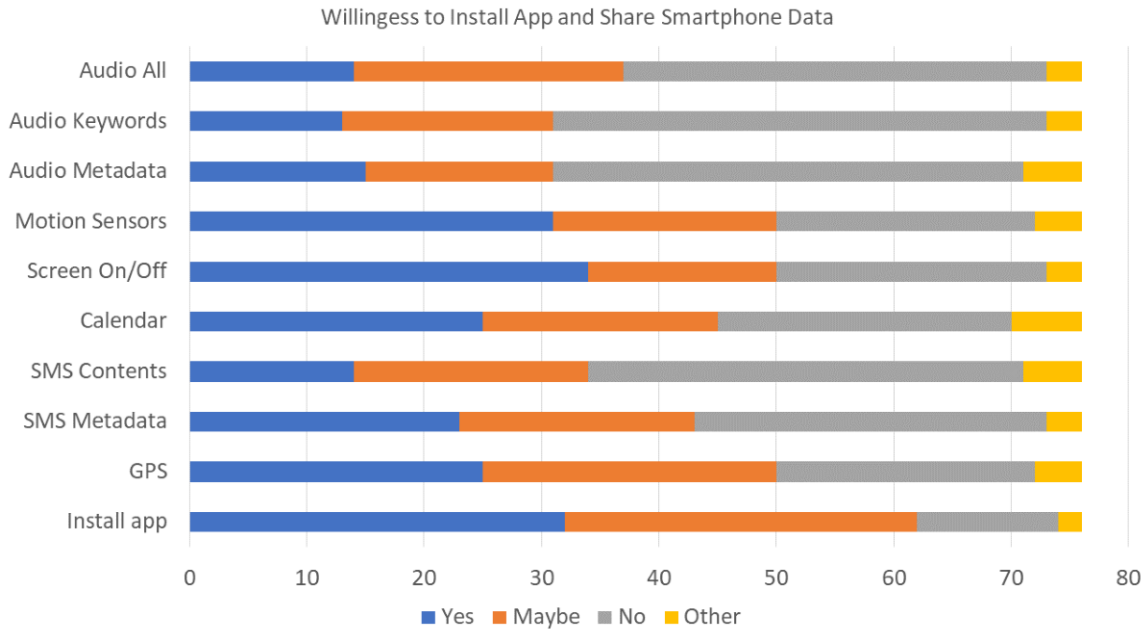


Figure 3.1: Respondent willingness to install application and grant permissions.

support, this is relatively good news that suggests that development solely on iOS and Android platforms is sufficient to support the vast majority of patients.

We observe that respondents' willingness to install a mental health monitoring application is generally positive, with 84% answering either yes or maybe when asked if they would install and use such an application on their smartphone. This finding suggests that there is a general positive interest in using smartphone applications to aid in mental health assessment. Considering that the questionnaire expresses that such a mental health monitoring application is believed to help both clinicians and patients, but makes no claims about proven effectiveness nor does it quantify said effectiveness in any way, it is plausible that if such applications are to be developed and proved to be effective then rates of adoption may be higher than reported in this study once patients are presented with these findings.

Respondent willingness to provide access to particular sources of data varies from a minimum of 43% answering yes or maybe in the case of audio metadata or keyword extractions (questions 10 and 11) to a maximum of 68% answering yes or maybe in the case of screen state (question 8). We feel that these results are encouraging considering the survey does not allay any potential fears that may be held by the respondents with regard to data security or data access. For example, in a production-ready application, data security is likely to be a considerable point of focus, with efforts to encrypt data in transit and at rest. If future applications incorporate these additional security features and effectively communicate them to users, willingness

could potentially increase in response. Furthermore, it may be possible to develop automated software algorithms to process much of the raw data and report higher-level statistics of interest to clinicians [70], [71], in a manner that removes direct human observation of patient data. This could serve to address patient fears that others would be scrutinizing them personally by, for example, reading their text messages or listening to their audio recordings. As our survey did not address any of these points, it may be possible that responses would be more positive given these guarantees.

It is worth noting that permission to collect data sources that are low resolution in terms of providing personal or private information, such as screen state (Q8) and motion sensor data (Q9), are the most likely to be granted. Contrast this with access to sources of data that offer much more insight into a person's private life, such as the contents of SMS messages (Q6) and unrestrained audio recording and analysis (Q12), which are among the least likely to be granted (only 20% and 19% yes responses, respectively). This could be interpreted as evidence toward the hypothesis that fear of scrutiny is responsible for unwillingness to provide access to data. If that hypothesis was confirmed, then this suggests that any automated analysis of data that removes humans from direct observation of the source data may be a method to improve patient willingness to provide data access. Further research is required to explore this hypothesis, however.

3.4.2 Comparison With Other Studies

The rates of smartphone ownership and use measured in this work is roughly in line with previous research in other areas. The results reported in this study are most similar to those measured in the United States, with ownership in the range of high 70% [72] to 80% [73]. The predominance of Android and iOS devices within the population under study is also in line with current market statistics [69]. The mobile phone ownership and use measured in this study was somewhat lower than the 96% rate measured by Zhang et al. [74] in a Chinese population.

While existing work demonstrates that smartphone and/or wearable-based health assessment tools are being adopted despite concerns around privacy [64], [65], the authors are unaware of any studies that have attempted to determine precisely which sources of data cause most concern. In the broader field of smartphone applications in general (i.e., without a focus on mental health care applications), Felt et al. [75] found that SMS messages were the data source that was most cause for concern, more so than Global Positioning System (GPS) location. While they did not consider audio recordings, the perception that SMS messages are more invasive of privacy than GPS

location is in line with our results. A similarly broad study into mobile phone user perceptions of privacy and security found that older users exhibited more privacy and security concerns [76]. This is in contrast with the population studied in this work, as we have measured a positive relationship between age and willingness to consent to audio recordings (i.e., older people are more willing to consent to audio collection).

3.4.3 Limitations

One limitation of the study is the small, concentrated population of respondents. All respondents presumably live within the Toronto area, and it is not clear how these results may generalize to the greater Canadian population or beyond. Another limitation of the study is the survey design. It is clear that respondents, in general, require more information to provide a sense of their willingness to provide access to data, as the proportion of maybe responses averaged across questions 4 through 12 was nearly one-third of respondents (28%). As mentioned earlier, 2 key pieces of information that would help respondents make more informed decisions are the effectiveness of the application in helping to monitor or manage their mental health if the data are provided and the risks involved and the steps being taken to protect the patient's privacy. Conveying both of these pieces of information to prospective application users will be a challenge for health care providers willing to employ mobile mental health applications.

Another fundamental limitation of this study is one inherent to surveys in general: it is not clear that survey respondents who expressed interest in the hypothetical application and willingness to consent to collection of their data through the application would actually consent in a real-life scenario involving real mental health applications. We surmise this would depend upon how effective the applications were shown to be and what patient perception of the risks would be.

Finally, it would be interesting for future work to determine if there were any measurable differences between how patients with different disorders might consent to access to their data. One interesting hypothesis to test would be whether patients with social anxiety disorder, who fear evaluation and scrutiny, would be less likely to provide data access, possibly due to the fear of observation or judgment from others characteristic of the disorder.

3.4.4 Conclusion

Given the potential for mobile technology to help patients monitor their mental health symptoms in a passive and pervasive way, it is helpful for researchers to understand

how patients may respond to requests for access to their personal data. General interest in such an application is moderate, with 41% of respondents indicating they would install a monitoring application and 43% of respondents indicating they may install such an application. Willingness to provide data collection across different sources ranges from 18% to 46%, with more intrusive or private sources of data being more likely to be withheld. Finally, we support previous findings [67], [68] that show mobile phone technology adoption alone will not pose a significant problem to fielding mental health applications, as 85% of respondents reported using an internet-connected smartphone daily.

Chapter 4

Study Design

This chapter describes the design of the study which was conducted to collect objective smartphone data and self-report measures of mental health from participants. The in-depth technical details of the smartphone data collection system are provided in Chapter 5. Results of the data collection, including number of participants and participant demographic data, are provided in Chapter 6.

4.1 Overview

An Android smartphone application was built, together with a centralized server system, to collect and store measurements of objective smartphone data. The types of data include samples of ambient audio, GPS location, screen state, light sensor data, and physical motion. Participants were recruited into a 2-week observational study where the application was run on their personal smartphone. Participants also completed self-report measures of social anxiety disorder, generalized anxiety disorder, depression, and functional impairment, in digital form through the study application. Participants were recruited through an online system, provided informed consent to engage in the study, and received $18.50 in compensation for completing the study. The study was approved by the University of Toronto Health Sciences Research Ethics Board, Protocol #36687.

4.2 Participants and Recruitment

Study participants were recruited through an online recruitment platform called Prolific [77]. Prolific is a company, and website, which allows researchers (both academic and market researchers) to recruit respondents for studies and tests, for a fee. It is similar to the well-known Amazon Mechanical Turk [78], but is more specifically

designed for scientific experiments [79]. Prolific maintains a list of registered participants and for each participant a variety of demographic data such as age, gender, primary language spoken, etc. When a study is offered to the participants in Prolific, the authors of the study specify inclusion criteria that limit which potential participants are allowed to enroll.

The following inclusion criteria were used for recruitment in this study:

1. Participants must be fluent in English
2. Participants must live in Canada
3. Participants must own and operate an Android smartphone daily, as their personal phone.
4. Participants must have completed at least 20 previous studies on Prolific
5. Participants must have completed 95% of their previous Prolific studies in a satisfactory manner (i.e., their participation was completed and approved).

No exclusion criteria were applied during recruitment.

4.3 Materials and Data

Three categories of data were collected from study participants. Demographic data of participants was provided by the Prolific platform on behalf of all consenting participants. The remaining two categories, objective smartphone-collected data and the self-report measures of mental health, were collected digitally through the study smartphone application.

4.3.1 Demographic Data

The following demographic data were collected:

1. Age
2. Gender
3. Employment Status
4. Nationality
5. First Language
6. Student Status
7. Country of Birth

4.3.2 Objective Smartphone-Collected Data

The smartphone application collected the following types of data:

1. **Location**: The application periodically collected GPS data in order to geolocate participants throughout the study. This data was collected once every five minutes.

2. **Battery**: The application periodically recorded participants' smartphone battery charge levels. This data was collected once every five minutes.

3. **Contacts**: The application recorded the number of contacts that participants call or send SMS (text messages) to, and the dates and times when these phone calls or messages occur. The names or numbers of contacts were not be recorded. The phone calls or SMS messages were not listened to, recorded, or read.

4. **Calendar**: The application recorded every time a participant created a calendar entry in their calendar app. The specifics of the calendar entries or events were not be recorded, only the date and time that calendar entries were created or modified.

5. **Screen State**: The application recorded every time that the smartphone screen turned on or off.

6. **Light Intensity**: The application used smartphone light sensors to periodically measure and record the intensity (brightness) of the ambient light in participants' surroundings. This data was collected once every ten minutes.

7. **Physical Activity**: The application periodically used smartphone motion sensors (including the accelerometer) to detect if participants are walking, running, in a car, or standing still. This data was collected once every five minutes.

8. **Audio**: The application used the smartphone microphone to periodically collect 15-second recordings of audio from participants' environments. This data was collected once every five minutes. These environmental audio recordings may pick up the sounds of participants, other individuals within sensing range of participants, or other sources of audio (e.g., television). These environmental audio recordings were processed by automated software in order to extract three sub-streams of data:

   (a) The average volume of each recording

   (b) The presence or absence of a speaking voices in each recording

(c) Spoken English-language words in each recording

A number of privacy-preserving considerations were made with respect to this particular data stream and are discussed in Section 4.5.

Greater technical detail on these collection mechanisms are provided in Chapter 5, where the smartphone-based data collection system is described in-depth.

4.3.3 Self-Report Measures of Mental Health

Participants completed four self-report measures in digital form within the smartphone study application at the beginning and end of the 14-day study. The measures were the following:

1. **Liebowitz Social Anxiety Scale (LSAS):** The LSAS is a 24-item scale used to assess the symptoms of social anxiety disorder. The properties of this instrument were discussed in Section 2.1.1.
2. **Generalized Anxiety Disorder 7-item scale (GAD-7):** The GAD-7 is a 7-item scale used to assess the symptoms of generalized anxiety disorder. The properties of this instrument were discussed in Section 2.1.2.
3. **Patient Health Questionnaire 8-item scale (PHQ-8):** The PHQ-8 is an 8-item scale used to assess the symptoms of depression. The properties of this instrument were discussed in Section 2.1.3.
4. **Sheehan Disability Scale (SDS):** The SDS is a 5-item scale used to assess general functional impairment. The properties of this instrument were discussed in Section 2.1.4.

Copies of the GAD-7 and PHQ-8, in their traditional paper format, are included in Appendix A (a sample of the LSAS can be viewed at [26]; a sample of the SDS can be viewed at [39]). Screenshots of the digitized versions of the GAD-7 and the PHQ-8, as they appeared in the study smartphone application, are included in Appendix D. It is worth noting that while these scales were initially developed and tested for pen-and-paper delivery, a review by Belisario et al. [80] found that, in general, self-administered survey scores do not differ when deployed by software application versus other delivery modes.

4.4 Study Procedure

Members of the Prolific community who met the inclusion criteria could read a description of the study, which included an informed consent guide. This study de-

scription is included in Appendix C. Those who consented to the study were then directed to a webpage that acted as the study entry point. This website directed participants to install the study smartphone application from the Google Play app store and provided them with login credentials for using the application. Once installed, the application guided participants through a short setup, where they were asked to provide the app with the necessary permissions to access their data, followed by a login. Immediately following setup and login, participants were asked to complete the set of four self-report measures in digital form within the study app. At this point, following the completion of the self-report measures, the application began to collect data in the background. No further actions or interactions with the study application were performed until the end of the study, exactly 14 days later, at the same time of day as the application installation. At this time, participants received a notification on their phone informing them that the study had ended and requesting that they complete the same set of four self-report measures done at the beginning, again in the application. Following completion of this task, participants were directed to uninstall the application from their phone and mark their study tasks as complete on the Prolific website. Subjects were then paid through Prolific's payment system.

4.5 Ethics and Privacy

The privacy and security of participants and their data was a key concern during the study, due to the intimate and highly personal nature of the data collected. One step taken to address this was to maintain participant anonymity as much as possible. The Prolific recruitment platform is designed in a manner to keep their users anonymous, and additional steps were taken to retain that anonymity throughout their participation in this study. All recruited participants were provided with randomized account logins for use in the study app, to prevent participants from using a user name which contains identifying information (e.g., their name or email address). Anonymity was also a concern during the handling and processing of participant data. Visualization of location data was done on modified map renderings which removed street names. A number of steps were taken to maintain anonymity while handling audio data, which will be discussed separately, below. Finally, to maintain the security of participants' data, all data was transmitted to study servers over encrypted channels and stored in encrypted form.

The study was approved by the University of Toronto Health Sciences Research Ethics Board, Protocol #36687.

4.5.1 Privacy-Preserving Actions for Audio Data

Audio recordings were processed entirely by automated software with no human intervention, and were deleted immediately upon processing to prevent them being listened to. These automated processing actions were the measurement of average volume, the detection of speaking voices, and the detection of English language words. Audio recordings were stored in encrypted form on participants' smartphones between the time at which they were produced and the time at which they were processed and destroyed, and in such a manner that they could not be unencrypted except by study staff. This additional action was performed in order to prevent the leaking of these recordings if a device were to be lost, stolen, or otherwise compromised. Additionally, the words detected from audio recordings were stored in randomized order to prevent recreation of speech. Further technical details on these privacy-preserving considerations are provided in Chapter 5, which discusses the technical design of the data collection and storage system used in the study.

Earlier versions of the study design intended to record and store these audio recordings indefinitely. Retaining access to the audio files, even after processing and extraction of relevant data, was desirable since it could enable further experimentation with novel processing techniques in the future. The ability to retain the audio files after processing was revoked by the research ethics board after it became clear that these audio recordings could potentially contain identifiable information of third parties without their knowledge, which is a breach of privacy. Canadian law has a one-party consent rule for the recording of private conversations, which means that as long as one member of a conversation consents to the recording (in this case, the study participant), then the recording is legal [81]. Consider, however, the case where a participant's device happens to record a conversation between third parties of which the participant is not a member (for example, sitting within earshot of two strangers conversing in a coffee shop). Such an audio recording would not be permissible.

After a series of discussions with the research ethics board and a privacy lawyer, the recording of participants' ambient audio was allowed to proceed if and only if the following conditions were met:

1. Audio recordings were encrypted such that participants, or any third parties who gained access to the audio recordings, could not listen to the recordings.

2. Audio recordings were only stored temporarily, until processing and extraction of relevant data, afterwards which they would be immediately deleted. This processing must be done in an automated fashion entirely by software, with no humans listening to the audio recordings.

3. The data extracted from the audio recordings must not contain personally identifiable information.

Conditions 1 and 2 above were both met by the careful engineering of the data collection system, to be described in Chapter 5. Condition 3 was met by limiting ourselves to only extracting data from the ambient audio which does not contain any such personally identifiable information. Ambient audio volume (sampled at the rate at which this study did — once every 5 minutes) and the presence or absence of speaking voices naturally do not contain personally identifiable information. The extraction of spoken words in ambient audio recordings was made to also comply with Condition 3 by randomizing the stored word order of the recognized words, such that transcripts of individuals' speech could not be recreated.

4.6 Summary

This chapter discussed the design of the study which was performed to collect objective smartphone data and measures of mental health from consenting participants. Participants were recruited anonymously from an online research platform, and all data collection and study activities were performed on smartphone software designed particularly for this study. In the next chapter, we describe the design of the smartphone software, and the supporting server and desktop software, which made this study possible.

Chapter 5

Data Collection System

This chapter outlines the software architecture of the system used to collect self-reported measures of mental health and objective smartphone-collected data from individuals.

5.1 System Overview

The software system used to collect and store study data from participants consists of four sub-systems: a study onboarding website, a smartphone application, backend server infrastructure including a database, and a data export application. Study participants first interacted with the onboarding website to gain access to the study smartphone application. The smartphone application, once set up and running, collected and transmitted data to a backend server system. The backend server system processed and stored data in a centralized database. This data could then be extracted from the database to a personal computer for analysis through use of the data export application. Figure 5.1 provides a high-level illustration of the overall system and flow of data. Details on each of the four sub-systems are provided in the sections that follow.

5.2 Onboarding Website

The onboarding website acted as the entry point to the study. Users of the Prolific recruitment website who read and consented to the terms of the study description on the Prolific website were immediately directed from Prolific to this onboarding website. The purpose of the onboarding website was to provide participants with all the necessary information and credentials to install and activate the study smartphone application.

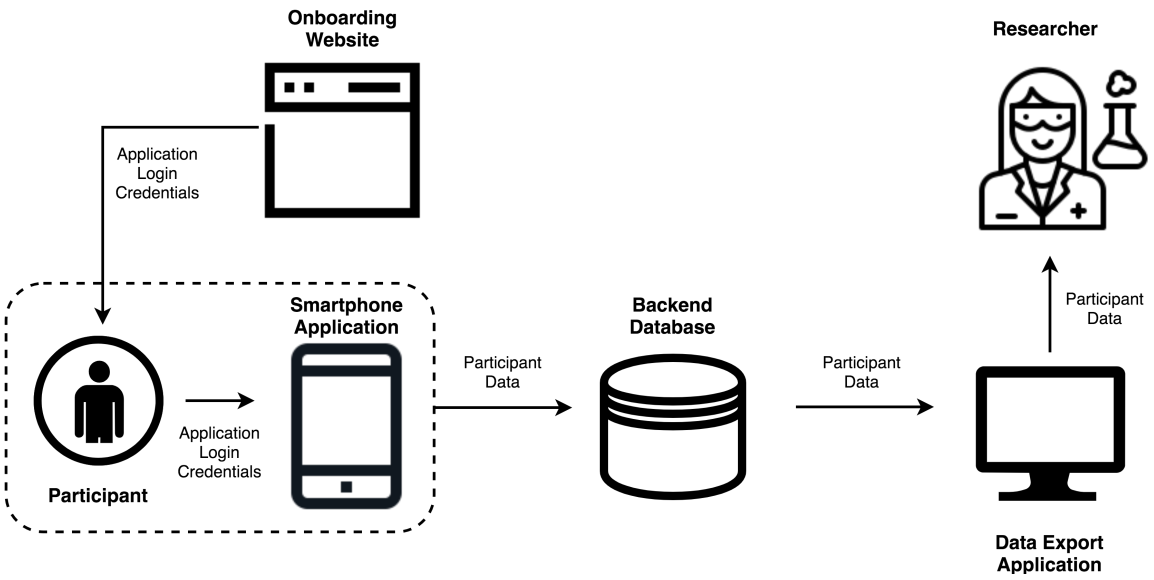


Figure 5.1: High-level architecture of the data collection system.

The onboarding website asked participants to begin their participation in the study by clicking a button to generate login credentials for use in the study smartphone application. These credentials were a randomly generated username and password. While these credentials were being generated, a user account was created and saved for each participant in the backend server. Additionally, a pair of encryption keys were also generated and stored in the backend database (the purpose of which will be described in Section 5.3). The randomly-generated username and password were then provided to the participant, who was asked to write them down and save them temporarily. Finally, participants were asked to click on an icon within the onboarding webpage which redirected them to the Google Play Store listing for the study application, where they could download and install the study application. Figure 5.2 provides an illustration of the flow of actions that are performed in participants' interactions with the onboarding website.

It should be noted that the onboarding website was intended for use on mobile phones in order to streamline the onboarding procedure. Having the onboarding website accessible from a mobile web browser meant that participants could interact with the onboarding website and install the study application without switching from a computer to their smartphone. It also made the study accessible to participants who only browse the web from their smartphone. While the onboarding website was fully functional when accessed from any web browser (desktop and laptop computers included), the study listing on Prolific requested that participants access the website using a mobile browser.

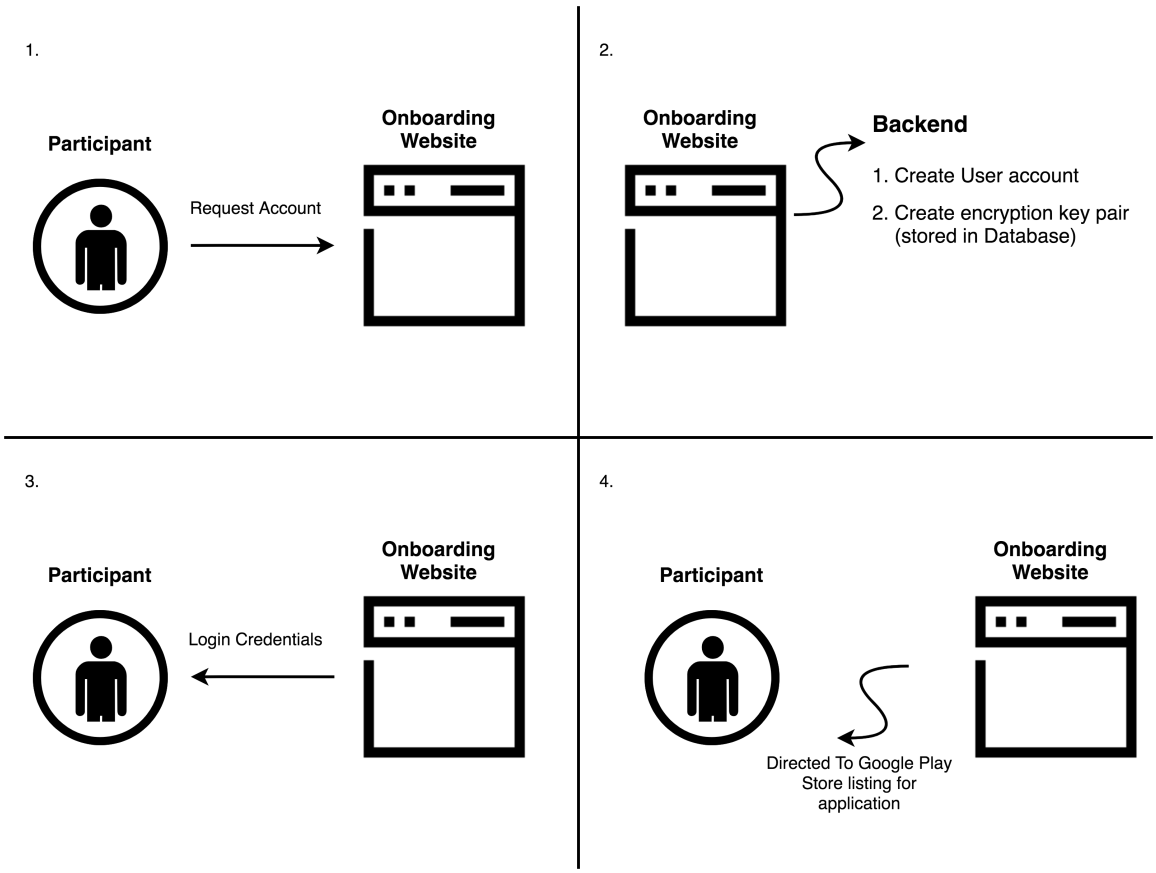


Figure 5.2: Action flow of onboarding website.

5.3 Study Smartphone Application

The purpose of the study smartphone application was to collect both the self-reported measures of mental health listed in Section 4.3.3 and the objective data listed in Section 4.3.2. The study smartphone application is an Android application called "Logger", available for free on the Google Play Store [82]. The source code for the smartphone application is comprised of 79 Java source files, containing a total of 7463 lines of code, not including libraries, APIs, or other external dependencies. The application was designed solely for Android, and not additionally Apple's iOS, because the iOS development model does not allow applications to passively collect audio and location data from the background without user interaction [83]. This was a key requirement for our framing of the research goals, so the decision to forgoe an iOS version of the application was made.

The study smartphone application consists of a number of functional components, the key component being that which collects all study data. Prior to any data collection, the application must be first authenticated and set-up. Immediately following

the setup procedure, participants completed the first set of self-report measures of mental health. Once completed, the application then began to passively collect objective data, and this data collection continued until the completion of the 14-day study. At that time, participants were once again asked to complete the same set of self-report measures of mental health, after which their participation in the study was concluded. The following three subsections provide details on these three components of the study application functionality.

5.3.1 Setup and Study Intake Functionality

Participants were required to complete a number actions within the study application upon first launch. Participants were presented with a guided setup procedure, with detailed instructions, in the application the first time it was launched from their smartphone. Participants were guided through the following actions:

1. **Disable Battery Optimizations**

   Participants were first asked to disable a setting on their Android smartphone, called battery optimizations, for the study application. This setting, which is enabled by default, severely impacts the ability of the application to passively collect data, and would significantly reduce the frequency of data collection. While disabling this feature can potentially decrease the battery life of the smartphone, brief testing showed no noticeable impact on battery life.

2. **Granting Permissions**

   In order for the application to be able to collect location, audio, calendar, and contact-based data, participants must grant the application explicit permission to do so. While participants were made aware of this data collection and had already given consent to do so by accepting the study on Prolific, the Android operating system also required this consent to be provided as part of its own permissions management model. Participants who denied any of these permissions were asked to withdraw from the study, and any further progress through the application was prevented. A screenshot of the application requesting the location data permission is shown in Figure 5.3a.

3. **Activating Location Data**

   Participants were asked to click a button to ensure that their smartphone's location data was enabled. While at this point in the setup they had granted permission to collect location data, it was still possible that their location settings

were turned off, which would have prevented collecting that data despite having permission to do so.

4. **Logging In**

   Participants were finally asked to log in to the study app by entering in the randomly-generated username and password which were previously given to them from the onboarding website.

5. **App Configuration**

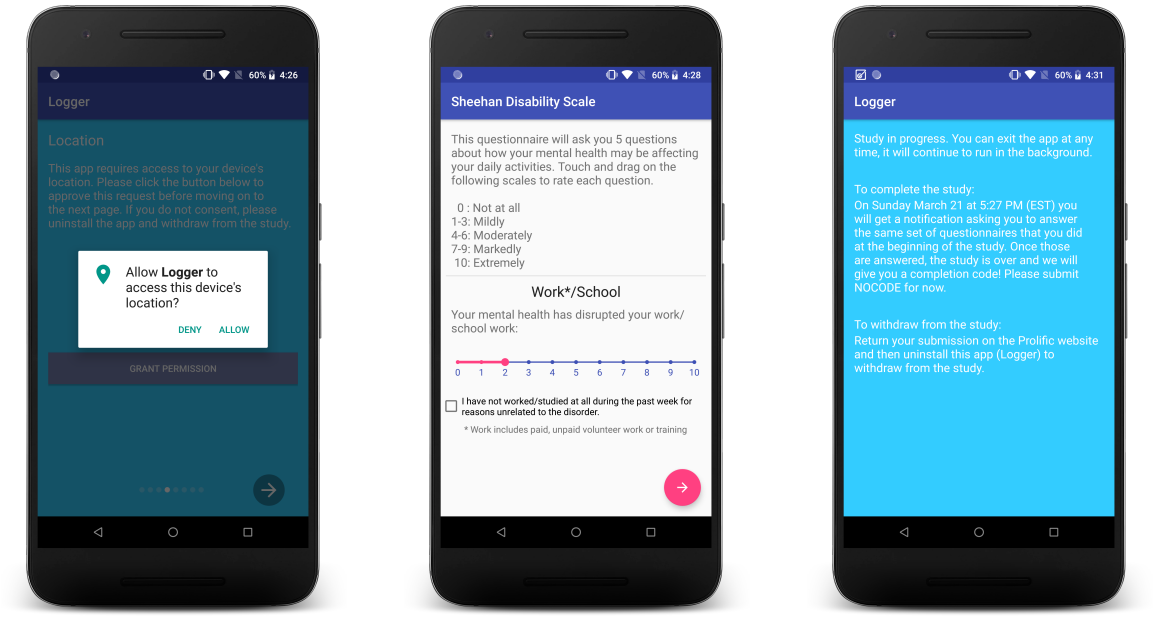
   Following login, the application automatically performed a set of operations, in the background, to configure it for passive data collection. An encryption key, generated for each participant at the time of their account creation, was retrieved from the backend server database and stored on their smartphone. An automated check was performed to ensure that the application could communicate with other parts of the backend server infrastructure. A time-based killswitch was also activated, which served to halt all passive data collection and to notify the participant that they must complete the final set of self-report measures at the conclusion of the 14-day study. This configuration process lasted only a few seconds, and participants were presented with a small loading screen for the duration. After this final step in the set-up procedure, the application was fully functional and participants were asked to complete the first set of self-report measures.

6. **Completion of Self-Report Measures**

   Participants were asked to complete the series of four self-report measures of mental health. These four measures —the LSAS, GAD-7, PHQ-8, and SDS —were presented to participants in digital form within the study application itself. A screenshot of the first question of SDS as it appears in the application is shown in Figure 5.3b. Following the completion of the self-report measures, participants were informed that no further actions are required for 14 days and that they could simply close the app. This message remained on the homescreen of the application for the duration of the study, and is shown in Figure 5.3c.

5.3.2 Passive Data Collection

The objective data listed in Section 4.3.2 will be reprised here, with additional technical details about the data and the mechanisms for sampling each data stream.



(a) Location permission request (b) Sheehan Disability Scale (c) Application homescreen

Figure 5.3: Three screenshots of the study smartphone application

Location, Battery, and Physical Activity

Data describing participants' location, physical activities, and charge level of their smartphone were all captured by the smartphone application through use of Google's Awareness API [84], which is a software library that allows application developers to gain easy access to contextual information about users of their software. The Awareness API recognizes physical activity states by collected and processing motion sensor data (accelerometer and gyroscope) in an automated fashion, without requiring users of the Awareness API to collect and process this raw motion sensor data themselves. A total of eight possible activities are detectable, including being in a vehicle, riding a bicycle, walking, running or being still (the other three activity labels represent some degree of indeterminate activity state). Location data was represented as participants' longitude and latitude on the Earth's surface. Battery data is the battery charge level of participants' smartphones. All three data were periodically sampled and recorded once every 5 minutes by the smartphone application.

Light

Ambient illuminance, measured in units of lux by the smartphone light sensor, was periodically sensed every 10 minutes by the smartphone application. This data was acquired through the `Sensor` class of the Android API [85].

Audio Recordings

The study smartphone application captured recordings of the environmental or ambient audio, using the smartphone microphone. The recordings were initiated periodically, every five minutes, and record 15 seconds of audio. These audio recordings were produced through use of the `MediaRecorder` class of the Android API [86]. Audio recordings were encoded in the Adaptive Multi-Rate Wideband (AMR-WB) encoding [87] and stored as 3GPP-format [88] files. These audio recordings were immediately encrypted (see Section 5.3.4 for a description of the encryption process) and remained on participants' devices until uploaded to the backend server file storage. Encrypted audio recordings were uploaded in baches, every 3 hours. This batched uploading was done to preserve battery life.

Three sub-streams of data were extracted from audio recordings:

1. Audio Volume: the average volume of each recording.
2. Voice Activity: the presence of absence of English-speaking voices in the recording.
3. Bigrams: English-language words detected in each recording. Each word is stored together with the word immediately following it, to form a pair of words or a *bigram*. For example, the transcript "what did you eat today" would yield the following sequence of bigrams: (what did), (did you), (you eat), (today ⟨EMPTY⟩). Note how contiguous bigrams in the sequence share a word; this is not a standard approach for bigram generation but was used in this work.

These three pieces of data were extracted from audio recordings by the backend server (see Section 5.4.3).

Calendar and Contacts

The study smartphone application recorded the creation of new contacts and new calendar entries created during the 14-day period of data collection. Contact data collection only captured the unique identifier assigned to each contact, not their name, phone number, or other personally-identifiable information. Calendar data collected contained the unique identifier of the calendar entry, and its start and end times. Calendar entry titles and descriptions were not captured in order to respect privacy.

Contact and calendar data were accessed through the `ContactsContract` [89] and `CalendarContract` [90] classes of the Android API, respectively. Neither of these classes provide an interface for the discovery of new contacts or calendar entries at the time of their creation. Instead, they simply act as wrappers around a database of

all existing entries at the time of querying. In order for the study application to detect only *newly created* entries, the application made an initial inventory of all existing entries at application set-up, and then periodically re-inventoried the Contacts and Calendar databases to look for new entries (i.e., entries that have not been seen since last inventory). Calendar entries were inventoried periodically, every three hours. Contact entries were inventoried every one hour. It must also be noted that the Calendar data accessed in this manner was only the data stored in the (default) Google Calendar application on participants' devices. If participants use another calendar provider and do not synchronize it with their Google Calendar, that data would not be accessible.

Screen State

The study smartphone application recorded the date and time of every instance of the smartphone screen turning on or off. This screen data is the only passively-collected data that is not collected periodically, but was collected in an event-driven manner in response to state changes (see Table 5.1). The application received a broadcast, through the `BroadcastReceiver` [91] class of the Android API, from the device's operating system when screen state changed. Note that, on newer Android devices, these broadcasts no longer specifically reflect screen state, but refer to when the device leaves and enters a responsive and interactive state (in other words, it tracks whether a device is "awake" versus "asleep") [92]. While this distinction is important to note, in practice it is not a great concern for our particular use case. In our use case, we wished to have some measure of how often individuals were on their device, and while "true" screen state is one indicator of device use, knowing when the device is asleep versus awake is also a good indicator, since devices will always be awake when being used, and will quickly enter a sleep state once not in use.

Data Stream	Sampling Period (minutes)
Location	5
Battery	5
Contacts	60
Calendar	180
Screen	N/A
Light	10
Physical Activity	5
Audio	5

Table 5.1: Sampling frequencies for the different streams of data collected by the study smartphone application

5.3.3 Study Exit Functionality

Precisely 14 days after entrance into the study, at the same time that the self-report measures were completed during application setup, participants received a notification on their smartphone asking them to re-complete the same four self-report measures of mental health. Following the completion of the self-report measures, they were directed to uninstall the study application and mark their study participation as complete on the Prolific website.

5.3.4 Audio File Encryption

Audio files produced by the study smartphone application were encrypted immediately after generation, and remained encrypted until being processing in the backend server. This encryption was done to preserve the privacy of both the study participants and any other individuals who may have been recorded speaking in the vicinity of the participants. Audio recordings were encrypted in a manner such that even participants themselves could not decrypt (and therefore listen to) the audio recordings.

This encryption scheme, where participants' devices automatically encrypt audio but have no ability to decrypt the audio is achieved through the use of asymmetric encryption (sometimes refered to as public-key cryptography) [93]. Unlike symmetric cryptography, where a single key is used to both encrypt and decrypt the data, asymmetric cryptography uses two keys (i.e., a *key pair*), referred to as the public key and private key. Data encrypted with one of the two keys can only be decrypted by the other. Either key can be used to perform the encryption, but conventionally the public key is used to perform the encryption and the private key is used to decrypt.

As each study participant proceeded through the study onboarding, a pair of 1024-bit RSA keys were generated. The key pair was saved in the backend database, and at application set-up only the public key was transmitted to the participant's device, while the private key remained in the database and was only accessible by the backend administrator. The RSA public key was not used to encrypt the audio directly, as data larger than the key size cannot be securely encrypted using RSA keys, and the audio files were significantly larger than 1024 bits.

To securely encrypt an audio file using the RSA public key, a one-time *symmetric* key was generated (128-bit AES key in CBC mode). One-time use keys generated in this fashion are often called *session keys*. The session key was used to encrypt the audio file, and then the session key itself was encrypted using the RSA public key. Note that this was easily done since the session key (128 bits) was smaller than the

RSA public key (1024 bits). The encrypted audio file and encrypted session key were then saved together as one secure file. The audio recording could only be decrypted by the session key, but the session key was also encrypted and could only be decrypted by the RSA private key. The RSA private key, however, is stored in the backend database. Note that, for each audio file, a new session key was generated and used, but not a new RSA public key. There was only one single RSA key pair generated for each study participant.

The generation of the session keys, encryption of the audio recordings, and the encryption of the session keys was all performed in the study smartphone application using base cryptographic functions from the `javax.crypto` package [94].

5.4 Backend Infrastructure

5.4.1 Overview

A backend server, sometimes referred to simply as a *backend*, is one or more computers which run software to support other software running on clients' computers (so-called *frontend* software). This distinction is common in web-based software, where the backend server is responsible for acting as a central store for clients' data and also for performing other important tasks which cannot safely be performed by the frontend software, such as user authentication. The advent of "cloud computing" has made it much easier for software developers to deploy software to backend servers without having to directly own, maintain, or manage any actual server computers. Many cloud computing vendors offer what is referred to as a *Backend-as-a-Service*, or BaaS, where the vendors maintain the physical server computer infrastructure themselves. Vendors then simply provide an easily-programmable interface for software developers to write and run backend server code on vendors' own machines.

The backend used by the data collection system was implemented using a BaaS offered by Google, called Firebase [95]. The database which stored all participant data, the file system which temporarily stored the encrypted audio, the server-side code which supported the onboarding website and the study smartphone application, and the onboarding website itself, were all built or otherwise enabled through a number of Firebase services. The remainder of this section describes the functionality of the Firebase backend and how it enabled the collection and storage of data throughout all the study procedures outlined earlier.

5.4.2 Support for Onboarding Website

The onboarding website described in Section 5.2 was itself hosted by Firebase's Hosting service [96]. The hosting service serves participants' web browsers the HTML, javascript, and other resources which comprise the onboarding website. The onboarding website, which was client-side software running in participants' web browsers, needed support from the backend in order to perform two key tasks (see Figure 5.2). The creation of each participant's user account and key pair for audio encryption must all be executed in the backend, since the backend offers a trusted compute environment that cannot be tampered with (something critical for secure generation of accounts and keys).

To do so, Firebase's Cloud Functions service was used [97]. Cloud Functions offer developers the ability to write server-side code, running on Google's own machines, which can be triggered for execution by events such as web requests from client-side code. Furthermore, this server-side code is managed by Google and automatically scales to increased demand without developers needing to request additional servers. This type of service enables the creation of what is often referred to as a *serverless* backend, since software developers do not need to interact with web server software when building a backend using such a service.

Two Firebase Cloud Functions were created to support the onboarding website. The first, which was triggered by a web request issued by the client-side website code, was used to create a user account for each participant within the Firebase Authentication service [98]. The second, which was triggered by the creation of the user account, was used to generate an RSA key pair and store it within Firebase's Cloud Firestore database [99]. This Cloud Firestore database was also used as the database for all data collected from participants.

5.4.3 Support for Smartphone Application

Firebase's Cloud Firestore database was used to store all participant data collected by the smartphone application during the study. Figure 5.4 illustrates the flow of data from the smartphone application to the Cloud Firestore database. All self-reported measures of mental health and all non-audio smartphone-collected data were directly written by the smartphone application to the database through the Cloud Firestore client library for Android [100].

The remaining data collected by the smartphone application — the encrypted audio recordings — were periodically uploaded from the smartphone application to Firebase's Cloud Storage [101] file storage system. Upon completion of every file

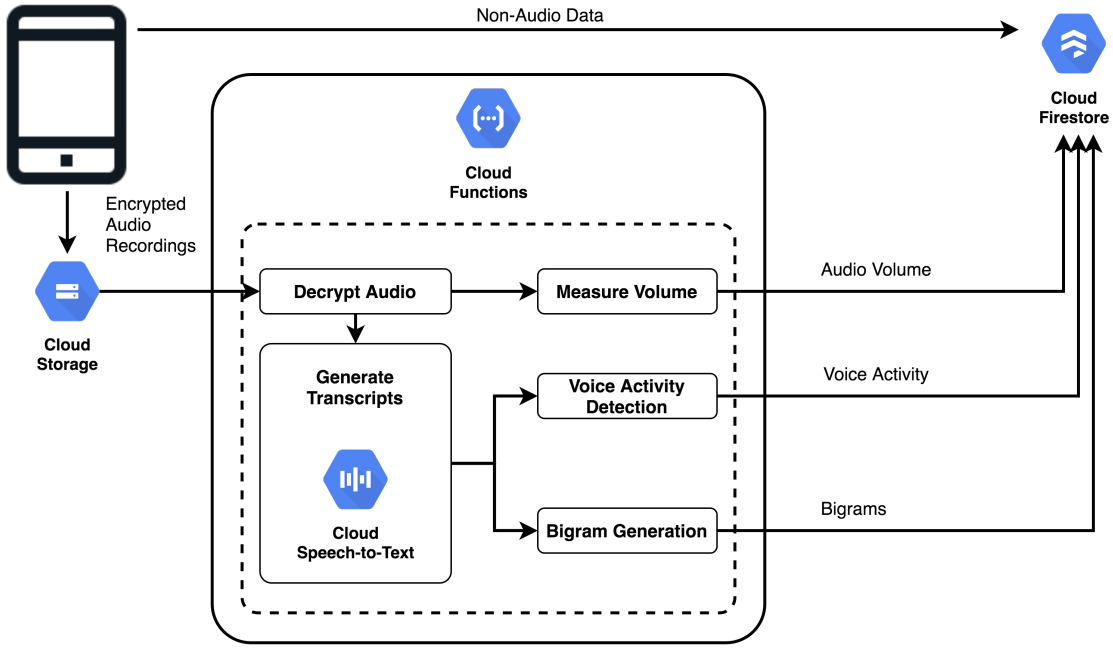


Figure 5.4: Architecture of the backend data storage system.

upload, the execution of a Cloud Function was triggered to extract the three streams of data from the audio recordings (audio volume, voice activity, and bigrams). Figure 5.4 gives a high-level illustration of this process.

The first operation performed by the audio-processing Cloud Function was to decrypt the audio file. The private key associated with the user which uploaded the file was read from the database and used to decrypt the session key bundled with the audio file. The decrypted session key was then used to decrypt the audio file. Once decrypted, the three pieces of data could then be extracted from the audio file.

The average volume of the audio file, in decibels, was measured using the FFmpeg audio processing library [102]. Prior to the detection of voice activity and bigram generation, a transcript of each audio file was generated using Google's Speech-to-Text automatic speech recognition software [103]. Once a transcript was generated, voice activity was detected by noting whether the transcript was empty or not. An audio file which produced a non-empty transcript was considered to contain voice activity. Bigrams, which are pairs of words consisting of each word in the transcript and the word immediately following it, were then also extracted from the transcript. Audio volume, voice activity, and bigrams extracted from each audio file were then saved in the Cloud Firestore database.

5.4.4 Database Schema

The Cloud Firestore database used to store all study data is not a traditional table-oriented or "relational" database. Firestore is a document-oriented database in which data is stored into *documents*, which together are grouped into *collections*. Each document is assigned a unique identifier, or *key*, to identify it within its associated collection. These keys can be automatically generated by the database or they can be assigned manually by the developer. Each document also contains a number of *fields*, which are used to store the desired data which comprise each document. These fields can be one of a number of different standard data types, such as strings, booleans, and numberic types.

A document structure was designed for each stream of smartphone-collected data (e.g., physical activity, average volume of the audio recordings, etc). Each document contains the sampled data of interest along with a timestamp indicating the data and time at which the data was sampled and a unique user id (UID) indicating which participant the data was collected from. Participant's UIDs are automatically generated by the Firebase Authentication system at account creation. All documents of a similar type are stored collectively in a single collection. All self-reported mental health measures were also stored in a collection of documents in a similar manner. Table 5.2 contains an illustration of how this data was organized into different collections of documents.

All documents of a similar type were stored together in a single collection, even those documents which contain data collected from different participants. Since documents contained a UID field which is unique for each participant, data from individual participants could be easily separated out by querying the database appropriately. Document keys were also always auto-generated, except for the documents which stored the RSA key pairs. Since only a single key pair was created for each participant, participants' UIDs were used as the keys to identify each RSA key pair in the database.

5.5 Data Export Application

The Data Export Application was designed to enable researchers to gain access to participants' study data for analysis. The Data Export Application is a standalone cross-platform application which can be run on any personal computer which has a Java Runtime Environment installed. The application makes use of the the Firebase Admin SDK [104] to read participant data from the Cloud Firestore database. Users of the Data Export Application must be able to first authenticate themselves as

Document Collection	Description	Document keys	Document fields	Field Data Type	Field Description
activity	Physical activity state	auto-generated	confidence	integer	Confidence of the given activity type
			timestamp	timestamp	Date and time of datum collection
			type	string	Activity type
			uid	string	User ID of participant
audioVolume	Volume of audio recordings	auto-generated	meanVolume	double	Average volume of recording
			timestamp	timestamp	Date and time of datum collection
			uid	string	User ID of participant
battery	Device battery charge level	auto-generated	batteryPercent	double	Battery charge
			timestamp	timestamp	Date and time of datum collection
			uid	string	User ID of participant
bigrams	Bigrams detected from audio recordings	auto-generated	first	string	First word in bigram
			second	string	Second word in bigram
			uid	string	User ID of participant
calendar	Calendar events	auto-generated	end	timestamp	Event end date and time
			id	integer	Unique event id
			start	timestamp	Event start date and time
			uid	string	User ID of participant
contacts	Contacts created	auto-generated	id	integer	Unique contact id
			uid	string	User ID of participant
light	Light sensor readings	auto-generated	illuminance	double	Illuminance of environment
			timestamp	timestamp	Date and time of datum collection
			uid	string	User ID of participant
location	Geolocation data	auto-generated	location	geopoint	Latitude and Longitude
			timestamp	timestamp	Date and time of datum collection
			uid	string	User ID of participant
privateKeys	RSA private keys for audio decryption	participant UID	privateKeyBase64	string	Base64-encoded private key
prolificUsers	Participants' Prolific information	auto-generated	prolificPid	string	Participant's Prolific ID
			sessionId	string	Prolific Session ID
			studyId	string	Prolific Study ID
			timestamp	timestamp	Date and time of datum collection
			uid	string	User ID of participant
publicKeys	RSA public keys for audio encryption	participant UID	publicKeyBase64	string	Base64-encoded public key
scales	Self-reported mental health measures	auto-generated	intakeOrExit	string	Which phase in study the measure was drawn
			results	integer array	Individual item responses
			scaleName	string	Name of self-report measure
			timestamp	timestamp	Date and time of datum collection
			uid	string	User ID of participant
screen	Screen state data	auto-generated	screenOn	boolean	State of screen
			timestamp	timestamp	Date and time of datum collection
			uid	string	User ID of participant
voiceActivity	Presence of speaking voices in audio	auto-generated	timestamp	timestamp	Date and time of datum collection
			uid	string	User ID of participant
			voiceActivity	boolean	Presence or absence of speaking voices

Table 5.2: Schema for all data stored in the backend database.

having administrator privileges within the backend system before data export can occur.

Participant data retrieved using the Data Export Application was stored locally on researchers' machines in comma-separated value (CSV) format. Each participant's data was stored in a single directory, with data from each data stream stored in a single CSV file within that participant-dedicated directory. While Firebase ensures that all data stored in both the Cloud Storage filesystem and Cloud Firestore database are encrypted and access-controlled, the CSV files produced by the Data Export Application are not encrypted. Users of the Data Export Application must therefore ensure that their personal computer used to store participant data has disk encryption capabilities and that this feature was enabled on their computer.

5.6 Conclusion

This chapter has described the various software systems built to enable the secure collection and storage of participants' mental health and objective smartphone-collected data. The design of a mobile website, Android smartphone application, Firebase backend, and Data Export Application was documented, with an emphasis on the

many features which establish and maintain security and reliability. The next chapter will describe the results of using this system in a study (described in Chapter 4) to collect self-reported measures of mental health and objective smartphone-collected data from study participants.

Chapter 6

Study Recruitment and Overview of Participant Data

This chapter discusses the deployment of the study (described in Chapter 4) to the Prolific online participant recruitment platform and provides a high-level summary of the data collected from study participants.

6.1 Participant Recruitment

From July 2019 to December 2019, 205 eligible Prolific members entered the study, with 112 participants completing all study tasks (details on participant withdrawals provided in Section 6.1.1). Participant recruitment was performed in a batched fashion, with multiple identical deployments of the study being performed one after the next, with only one deployment active at a time. At the conclusion of one study deployment, participants were paid, data was reviewed for integrity, and the next deployment was initiated. Table 6.1 provides information on the six deployments of the study, including the intended number of participants to be recruited, the number of Prolific users eligible to enter the study at that time of deployment, and the number of participants who entered, withdrew from, failed to complete, and completed the study.

The study was conducted in a batched fashion, as opposed to simply recruiting all participants at once, since it was unclear how the entire system and study procedure would scale to a large number of concurrent participants. While we were confident that the software would adapt to meet the load imposed by a large number of users, and it was indeed designed to do so, there remained specific challenges related to the monitoring of the study that drove us to limit the number of concurrent study participants.

Deployment	Launch Date	Number of Participants					
		Intended	Eligible	Entered	Withdrew	Incomplete	Completed
1	10-Jul-19	10	400	21	11	0	10
2	03-Sep-19	19	406	35	16	1	18
3	02-Oct-19	20	402	27	7	1	19
4	17-Oct-19	22	386	43	21	1	21
5	04-Nov-19	30	367	52	22	3	27
6	25-Nov-19	30	329	27	9	1	17
Total				205	86	7	112

Table 6.1: Summary of participant recruitment trials

The first monitoring challenge was that participants could have difficulty completing the application setup or other required tasks and may request help from the study organizer. The Prolific recruitment platform offers participants the ability to anonymously message study organizers for help or clarification with study tasks, and the assistance of participants in a real-time fashion is difficult when potentially multiple participants are requesting help. Secondly, independent of participants requesting real-time help, monitoring was still deemed necessary to ensure the correct setup of the application on participants' own smartphones. Significant testing of the smartphone application was performed prior to deployment, but the sheer number of Android device vendors and models is such that it was infeasible to test the application on all devices. In order to ensure that the study application was running correctly across participants and their devices, the backend system was monitored and key log messages and data were inspected to catch any potential problems during study intake.

In practice, few messaging interactions with participants were encountered, and the automated testing of the application during setup appeared to catch most cases of software problems without any monitoring. In total, across all deployments, monitoring of the login and setup procedure resulted in messaging interactions with nine participants. Four participants were messaged because they had accepted the study but failed to login in to the system; these participants later withdrew from the study without responding. Three participants failed to successfully disable battery optimizations for the study application; these participants withdrew from the study. One participants' application failed the automated set up, and withdrew from the study as a result. Finally, one participant could not successfully login to the application, and was also forced to withdraw from the study.

After a first small deployment to 10 participants, which proceeded smoothly, the next deployment was run on a larger group of 20 participants. As we gained confidence in the reliability of the software and the study design, the sizes of the deployments increased to a final size of 30. No further study deployments were made after the

sixth because it became clear that we had exhausted the pool of eligible participants. Participation was limited to Prolific users residing in Canada, of which there were only roughly 450 total in the year 2019. In all deployments prior to the last, all free slots in the study were quickly filled within a number of hours of deployment to the Prolific website. The sixth deployment, however, was left open to entry for multiple days and still failed to fill completely, leading us to believe that the pool of eligible *and* willing participants had been exhausted.

6.1.1 Participant Withdrawals

Withdrawals from the study were common, with 86 of 205 (42%) participants choosing to withdraw at some point in the study. The majority of withdrawals, 87% (75/86), occurred before a successful log-in to the study application (see Figure 6.1 for an illustration of the withdrawals throughout the timeline of study actions). Having observed how most withdrawals occurred early in the study, we provide two possible hypotheses for the high number of withdrawals. The first is the relative difficulty in setting up the study app on a personal smartphone, which is more complex than the typical task asked of subjects recruited on the Prolific platform. The setup procedure included turning off the smartphone's battery optimizations for the study application, which requires some facility with Android settings.

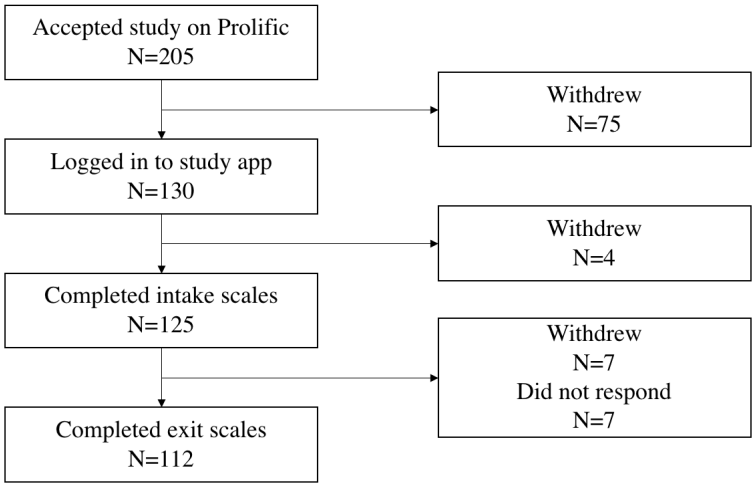


Figure 6.1: Flow chart of study recruitment and participant withdrawals.

The second is the possibility that subjects were unwilling to provide the app with the permissions necessary for data collection. Recall that the study application asks users to grant permissions before log-in, which might explain why 87% (75/86) of the withdrawals occurred without logging in. Despite the fact that the participants had already agreed to the terms of the study (which describe the data collection performed

by the application), it is possible that once explicitly confronted with the notion of such intimate data collection, participants chose to withdraw. It is worthwhile to note that the number of withdrawals (42%) is roughly in line with the results of our study described in Chapter 3, in which a combined 59% of respondents indicated either partial or no willingness to install a smartphone application to monitor their mental health.

6.2 Participant Demographics

At the completion of every study deployment, Prolific provides researchers with demographic data on all participants who successfully completed the study tasks (no demographic data is provided for participants who withdraw). This demographic data includes participant age, gender, student status, and employment status. Participants' demographic data is summarized in Table 6.2.

Age	n	(%)
18 to 24	32	(29%)
25 to 34	51	(46%)
35 to 44	18	(16%)
45 to 54	7	(6%)
55 to 64	4	(4%)
Gender		
Female	51	(46%)
Male	61	(54%)
Student		
Yes	39	(35%)
No	73	(65%)
Employment		
Full-Time	68	(61%)
Part-Time	19	(17%)
Unemployed	14	(13%)
Other	6	(5%)
Not in paid work	4	(4%)
Due to start a new job	1	(1%)

Table 6.2: Summary of participant demographic data

The sample of participants was quite young, with an average age of 30.6 years (SD 9.4). Three quarters of participants were under the age of 35 years old. Participants were 46% female, indicating a slight imbalance in gender representation. A sizeable minority (35%) of participants reported being students. Finally, 61% of participants

reported being in full-time employment, 17% of participants were employed part-time and 13% of participants reported being unemployed. Note that the "Not in paid work" state for employment includes, for example, homemakers, retirees, or those who cannot work due to a disability.

6.3 Self-Report Measures of Mental Health

At the completion of the study deployments, participants' responses to the four self-report measures of mental health, completed at study intake and study exit, were scored. Table 6.3 presents the mean and standard deviation of the measures taken at intake and exit. Additionally, in order to test whether statistically significant differences in mean scores exist between the administration of the measures at intake and exit, a paired samples t-test was performed for each measure. Mean scores on all four measures were lower at exit than at intake, and mean scores on the GAD-7 and SDS were significantly lower (at a 5% significance level). Figure 6.2 also provides histograms which illustrate the distributions of participants scores, taken at study exit, on the LSAS, GAD-7, PHQ-8, and SDS.

Measure	Mean Score (SD)		t-statistic	P
	Intake	Exit		
LSAS	58.0 (27.5)	55.8 (25.5)	1.82	0.071
GAD-7	7.7 (5.3)	6.7 (4.4)	3.22	0.002
PHQ-8	9.1 (5.5)	8.6 (5.3)	1.53	0.129
SDS	12.6 (8.3)	11.1 (7.6)	2.61	0.010

Table 6.3: Comparison of mean scores on all self report measures collected at study intake and study exit

To gain a sense of the severity of these scores, it is worthwhile to compare scores on the LSAS, GAD-7, and PHQ-8 to their respective diagnostic thresholds for screening for social anxiety disorder, generalized anxiety disorder, and depression, respectively. Table 6.4 shows the results of this screening operation.

Measure	Diagnostic Threshold	Disorder	Participants above threshold (%)	
			At Study Intake	At Study Exit
LSAS	60	Social Anxiety	43 (38%)	44 (39%)
GAD-7	10	Generalized Anxiety	36 (32%)	27 (24%)
PHQ-8	10	Depression	47 (42%)	40 (36%)

Table 6.4: Results of using the LSAS, GAD-7, and PHQ-8 measures to screen participants for social anxiety disorder, generalized anxiety disorder, and depression, respectively

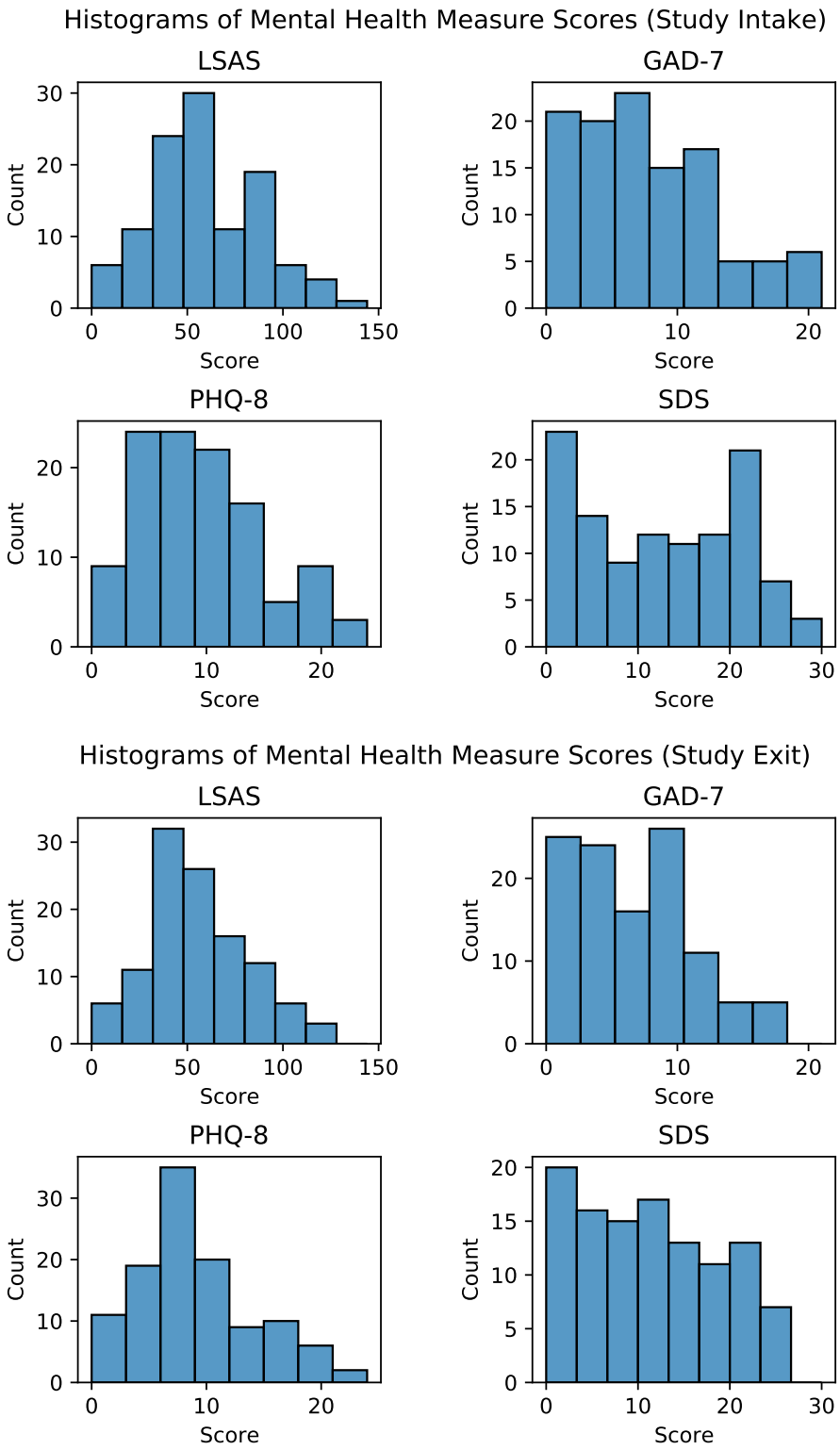


Figure 6.2: Histograms of participant scores on the LSAS, GAD-7, PHQ-8, and SDS measures

These results reveal that the sample of participants, despite being recruited from a healthy population, had a high prevalence of depression and anxiety. Data reported by the Government of Canada in 2006 estimate a 12-month prevalence of major depressive disorder at 4.8% [18], compared with positive screening rates of 42% (at study intake) and 36% (at study exit) in our sample. The same 2006 report lists a 12-year prevalence of the combined class of anxiety disorders at 4.8%, a figure that is significantly lower than the positive screening rates for generalized anxiety disorder (32% at study intake and 24% at study exit) and social anxiety disorder (38% at study intake and 39% at study exit) observed in our sample. Given that the thresholds used for screening were all shown to have high specificity (and therefore low false positive rate), it seems unlikely that these high rates were solely a result of false-positive screenings. Instead, 2 possible explanations are proposed, which may be jointly responsible. First, the Canadian population of subjects on the Prolific recruitment platform may have elevated rates of mood and anxiety disorders with respect to the general Canadian population. It may be that individuals who choose to find work on the platform do so to supplement their income. Individuals suffering from mental health conditions may have difficulty finding high-paying jobs because of their condition. Second, that subject sampling was impacted by self-selection bias. In other words, Prolific participants who struggle with mental health to some degree may be more likely to have chosen to enroll and remain in a study that focuses on mental health.

6.4 Objective Smartphone-Collected Data

This section provides a brief initial summary of the objective smartphone data collected from study participants. The emphasis here is on investigating how much data was collected and to determine the effective sampling rates achieved by the smartphone application for each of the data streams which it collects. Chapter 7 provides visualizations of the individual data streams and in-depth analyses of the data.

Table 6.5 provides a high-level summary of the amount of data collected, in each data stream, from study participants. The *Participants with Zero Samples* column lists the number of participants from which no data samples were collected. The *Samples per Participant* column lists the median number of samples collected per participant, and also provides the inter-quartile range (IQR) as a measure of the dispersion of these values (the interquartile range lists the 25th and 75th percentiles, between which the middle 50% of values lie). The *Sampling Period* column compares the nominal sampling rate at which the smartphone application was programmed to

sample data streams (for those data streams that are sampled periodically) to the median *observed* sampling rate, which is computed as the median of participants' average sampling periods. Each individual participant's average sampling period was computed as the total duration of time during which data was collected divided by the number of data samples collected in that time period. The inter-quartile range is also provided for the observed average sampling periods. Participants which yielded no samples where excluded from the computations of Samples per Participant and Sampling Period.

Data Stream	Participants With Zero Samples: n (%)	Samples Per Participant: Median (IQR)	Sampling Period (mins)	
			Nominal	Observed: Median (IQR)
Activity	2 (2%)	3955 (2465 - 4018)	5	5.1 (5.0 - 6.5)
Audio Volume	8 (7%)	3880 (3182 - 3972)	5	5.2 (5.1 - 6.4)
Battery	1 (1%)	4025 (2662 - 4066)	5	5.0 (5.0 - 7.5)
Bigrams	11 (10%)	3364 (1747 - 5418)	-	-
Calendar	112 (100%)	-	-	-
Contacts	112 (100%)	-	-	-
Light	5 (4%)	2015 (1338 - 2096)	10	10.0 (9.4 - 11.1)
Location	7 (6%)	3807 (1209 - 4023)	5	5.4 (5.1 - 13.0)
Screen	2 (2%)	1737 (655 - 2755)	-	-
Voice Activity	8 (7%)	3880 (3182 - 3972)	5	5.2 (5.1 - 6.4)

Table 6.5: Aggregate number of samples collected and sample period for each data stream collected from study participants

Focusing upon the Participants with No Samples column of Table 6.5, it is clear that no calendar or contact data were collected from any study participants. This could be due to bugs (errors) in the study application, but this seems unlikely to result in no data collection from all participants, since both the calendar and contact data collection features were tested and functional prior to deployment. Another possibility is simply that no participants added no new contacts to their device or calendar entries to their Google calendar during the time in which they were enrolled in the study, which is plausible considering the study was only 14 days long. Recall that that the calendar data collection only discovers and records the creation of calendar entries in the Google calendar smartphone application, and that it is not clear how many individuals use the Google calendar application specifically. A second observation which can be made is that, excluding than Calendar and Contact data streams, the number of participants with data that is entirely missing (i.e., not a single data point was collected) is low, being equal to or less than 10% in all other data streams. Note, however, that these rates do not count participants which were missing some (but not all) data, but only those who were missing all data entirely.

Shifting focus to the Samples per Participant column of Table 6.5, one can make the observation that, for all data streams other than calendar and contacts, there was a large number of data points collected (on aggregate). The median number of data points collected per participant ranges from 2015 (light sensor data) to 4025 (bigrams — pairs of spoken words detected in ambient audio recordings). This is to be expected, however, given the duration of the study and the sampling periods for the data streams.

The Sampling Period column of Table 6.5 allows us to assess how close the application came, in practice, to achieving its nominal sampling rates. Data streams which were intended to have a sampling period of 5 minutes were sampled, on average, slightly more slowly. The largest deviation between nominal and observed sampling periods occurred in the case of Location data, where the median average sampling period was 5.4 minutes. Looking at the interquartile ranges, however, shows a different picture. The upper value in the interquartile range corresponds to the 75th percentile of the data, where 25% of participants had average sampling periods larger than this value. Focusing on the upper range of the IQR, we can now see that some participants' data was sampled with a much lower effective frequency than intended. The 75th percentile of the average sampling periods for the Location data stream was 13.0 minutes. This indicates that 25% of participants had their location data sampled with an average sampling period greater than 13.0 minutes, which is quite a large discrepancy between nominal and observed sampling rates for these participants.

6.4.1 Challenges with Passive Data Collection

The discrepancy between the nominal and effective sampling rates is an important finding to discuss. This result, while undesirable, is not surprising given the trend of smartphone vendors trying to improve battery life while smartphones continue to grow more powerful and capable. The Android operating system introduced a key battery life-preserving feature, called Doze mode [105], in Android version 6. In an attempt to preserve battery life, this feature imposes restrictions on how often applications can perform actions in the background, that is, when users are not interacting with the application. Given that all data-sampling performed by the study application occurs in the background, and the user does not interact with the application except during setup and exit, Doze mode directly targets applications such as the one designed for this study.

In short, even when users opt-in to disabling battery optimizations for a specific application (as was the case with the study application during setup), Doze mode still prevents an application from waking a device from a sleep state to perform

background operations more than once every 9 minutes. Devices enter a sleep state, which is a very low-power state, only when users are not interacting with them. When devices are not in a sleep state, background actions can happen more frequently. This explain why, for example, a sampling period shorter than 9 minutes can be achieved (see Table 6.5).

In newer versions of Android (Android 9 and above), another set of restrictions are also imposed which further limit applications' abilities to regularly perform background operations when users do not regularly interact with those applications. Infrequently used apps get placed into low-priority App Standby Buckets [106], allowing the Android operating system to further limit those applications' abilities to collect data in the background at high frequencies. This explains why data sampling which is intended to be slower than the 9-minute period limit imposed by Doze is still further throttled to realize even slower effective sampling rates.

The so-called "war on background processing" shows no signs of slowing, and restrictions are likely to become even more severe in the future, which has significant implications for this type of research. Parties interested in performing high-frequency background data collection on Android smartphones may be forced to resort to collecting data using a paired smartwatch or other wearable device. Many of these devices are designed for frequent, passive data collection, and the data which they collect can still be accessed by a companion application running on a paired smartphone or through a web interface (see, for example, Amazon's Halo device [107]).

6.5 Conclusion

This chapter has described the deployment of the study (described in Chapter 4) to the Prolific online recruitment platform. The participant recruitment procedure was explained, and data was provided on the number of participants who entered, withdrew from, and completed the study. An overview of the completed participants' demographics was also provided, along with summaries of their self-reported measures of mental health. Finally, a high-level overview of the amount of smartphone-collected objective data was provided, along with a discussion of some of the challenges to passive data collection on Android devices. The next chapter is the first of three chapters which provide an analysis of participants' objective smartphone-collected data and detail the design of features which capture aspects of participants' mental health.

Chapter 7

Activity and Location-Based Features

This chapter begins the analysis of the smartphone-collected data, in which features were designed to detect mental health signals from each of the data streams collected. This chapter focuses upon the the design and extraction of features derived from participants' activity recognition (described in Section 7.1) and GPS location data (described in Section 7.2). The features presented in this chapter are a mix of novel features and features which replicate prior work; replicated features will cite the prior work in which they were first designed and presented.

7.1 Activity Recognition Data

7.1.1 Data Overview

Activity recognition data, provided by the Google Awareness API [84], detects and categorizes which activity an individual is engaged in at the time of sampling. Activity categories include running, walking, standing still, being in a vehicle, cycling, or an unknown activity. The smartphone application was designed to sample this data at a nominal rate of once every 5 minutes, but the sampling period achieved by the application when deployed to some participants' devices was, in effect, much longer (readers may refer back to Section 6.4.1 for a discussion on the discrepancy between nominal and effective sampling rates of the study smartphone application). Since missing data presents a difficulty when attempting to infer any patterns in the data, study participants with too few data points were excluded from analysis and feature generation.

The minimum number of data points required for a subject to be included in

analysis was set at *half* the number of data samples which would be expected if the application sampled activity recognition data at the nominal rate. A sampling period of 5 minutes over the 14-day study would yield 4,032 samples, and so participants which had 2,016 or more activity recognition samples collected by the study application were included in the analysis. Applying this threshold for inclusion resulted in a subset of 83 participants for analysis (of 112 total participants).

7.1.2 Analysis and Feature Design

To gain an understanding of the amount of time spent in each activity state by participants, the percentage of total activity recognition samples which detected each activity recognition category were computed for each participant. For example, for a specific participant, 0.5% of activity recognition samples detected that the participant was in a vehicle, 0.1% of samples detected them on a bicycle, 3.0% of samples detected them being on foot, 90.3% of samples detected them being still, and 6.1% of samples detected an unknown activity.

Table 7.1 presents high-level summary statistics regarding these activity state percentages across participants. For each recognized activity state, the table lists the mean percentage of activity recognition samples which fell into that category, the minimum, the first, second, and third quartile, and the max (i.e., the fourth quartile).

Activity	Percentage of Total Activity Recognition Samples					
	Mean (SD)	Min.	First Quartile	Second Quartile	Third Quartile	Max.
In Vehicle	1.9 (1.9)	0.0	0.6	1.3	2.8	9.6
On Bicycle	0.1 (0.2)	0.0	0.0	0.1	0.1	1.0
On Foot	1.5 (1.3)	0.0	0.6	1.1	2.1	7.0
Still	91.3 (5.3)	75.0	89.5	92.0	95.0	99.4
Unknown	5.2 (4.0)	0.4	2.7	4.1	6.2	21.6

Table 7.1: Statistics on physical activity rates of study participants (n=83).

From the data shown in Table 7.1, it is clear that study participants' devices are mostly still and not moving. On average, 91.3% of activity recognition samples detected individuals' devices as being still, with 1.9% of samples detecting devices as being in a vehicle, 0.1% of samples detecting devices as being on a bicycle, 1.5% of samples detecting devices as being on foot, with a final 5.2% of samples detecting an unknown activity. We intentionally refer to the activity recognition samples as detecting the activities of *devices* — and not *individuals* — as it is possible that individuals could be engaging in physical activities without having their device on their person (in which case the activity recognition samples would detect a "Still" activity).

Activity Rate Feature

Given that fatigue and low energy are defining characteristics of depression [17], a feature which provides some measure or indication of an individual's activity levels is likely to be (negatively) correlated with depressive symptoms. While neither generalized anxiety nor social anxiety are characterized directly by low physical activity, low activity might indicate that an individual is avoiding interaction with others, which can be indicative of social anxiety [17]. For these reasons, the activity recognition data is used to build a feature which quantifies physical activity.

The activity recognition data provides a quantification of individuals' physical activity across five classes of activities. To create a single, univariate metric of physical activity, the *Activity Rate* feature was created. The Activity Rate of a participant is computed as the percentage of that individual's activity recognition samples which correspond to any activity class other than "Still" (i.e., In Vehicle, On Bicycle, On Foot, or Unknown):

$$\text{Activity Rate} = \frac{|\text{AR}_{activity!=Still}|}{|\text{AR}|} \times 100\% \tag{7.1}$$

where AR represents the set of activity recognition samples belonging to a participant.

Summary of Measured Feature Values

Table 7.2 provides descriptive statistics on the distribution of feature values computed for the sample of the 83 participants with sufficient activity recognition data. Participants were, on average, in an active state for 8.7% of the time throughout the study. The minimum activity rate achieved by any participant was 0.6%, and the maximum activity rate achieved by any participant was 25.0%.

Feature	Descriptive Statistics					
	Mean (SD)	Min.	First Quartile	Second Quartile	Third Quartile	Max.
Activity Rate	8.7 (5.3)	0.6	5.0	8.0	10.5	25.0

Table 7.2: Descriptive statistics of the Activity Rate feature (n = 83 participants).

7.1.3 Correlations with Mental Health Measures

To understand how this feature captures aspects of mental health, Pearson correlations were computed between participants' Activity Rate feature and their four

self-report measures of mental health. The correlations are shown in Table 7.3. Activity Rate was most strongly correlated with depression, as measured by the PHQ-8 (r = -0.26, P = 0.02).

Feature	Correlations with Mental Health Measures: r (P)			
	LSAS	GAD-7	PHQ-8	SDS
Activity Rate	-0.24 (0.03)	-0.07 (0.53)	-0.26 (0.02)	-0.15 (0.17)

Table 7.3: Correlations between the Activity Rate feature and mental health measures (n = 83 participants).

7.1.4 Discussion

Aligning with our hypothesis, a higher Activity Rate is associated with lower symptoms of depression (r = -0.26, P = 0.02). Activity Rate is also negatively correlated with symptoms of social anxiety (r = -0.24, P = 0.03), which suggests that the activity rate feature may be capturing some degree of avoidance behavior. Correlations with symptoms of generalized anxiety are, by contrast, much weaker (r = -0.07, P = 0.53). This is unsurprising given that avoidance behavior is much more a hallmark characteristic of social anxiety than it is of generalized anxiety. Finally, Activity Rate is also negatively correlated with general impairment as measured by the SDS (r = -0.15, P = 0.17).

7.2 Location Data

7.2.1 Data Overview

The location of participants was periodically measured in latitude and longitude by their smartphone's GPS sensor. This data was sampled by the smartphone application at a nominal rate of once every 5 minutes, but the effective sampling period achieved by the application was, on aggregate, longer than this (the 75th percentile of the average sampling periods of location data was 13.0 minutes). As with all data streams, a threshold on the minimum number of samples per participant for inclusion in analysis was set in order to handle missing data. This threshold is similar to the threshold used in the analysis of all the periodically-sampled data; participants are required to have half the nominal number of samples expected (2,016 samples). Applying this threshold resulted in a sample size of 71 participants for analysis.

7.2.2 Analysis and Feature Design

Two hypotheses drove the creation of features derived from location data. The first, focusing upon depressive symptoms, is that individuals who experienced stronger depressive symptoms would exhibit less activity as inferred from GPS location data, than healthier individuals. This hypothesis is consistent with some of the diagnostic criteria for depressive episodes. Namely, a loss of interest in activities, slower movement, and a decrease in energy [17].

The second hypothesis, related to anxiety disorders, is that the avoidance behavior associated with the disorders would also manifest itself in an observable way in individuals' history of locations visited over time. While this is less straight forward to measure than physical activity, avoidance of certain situations may be measured by proxy. Consider, for example, individuals which leave the home very infrequently, or those who do leave the home but only travel to the same few number of locations. This less dynamic activity could serve as a proxy measure for avoidance behavior, whether it be avoidance of locations which are likely to place the individual under the scrutiny of others (in the case of social anxiety), or avoidance of locations which are more general triggers for anxiety (in the case of generalized anxiety disorder).

With these two general hypotheses in mind, the engineering of features derived from participants' GPS location data will now be presented. The first eight features presented were all designed and first described by Saeb et al in [44]. Two novel features described at the end of this subsection, *Exits from the Home* and *Home Stay Entropy*, take inspiration from that work.

Location Variance **Feature**

The Location Variance feature acts as one measure of the range over which participants travel. This feature was developed by Saeb et al. in [44], and is computed as the logarithm of the sum of the variances in latitude and longitude of the GPS location samples:

$$\text{Location Variance} = log(\sigma_{lat}^2 + \sigma_{lon}^2) \tag{7.2}$$

where σ_{lat}^2 and σ_{lat}^2 are the variance of the latitude and longitude of participants' location samples, respectively.

Number of Locations **Feature**

The Number of Locations feature is intended to count the number of unique areas visited by participants. Given the large number of GPS location samples captured,

and the fact that GPS location can vary slightly even when an individual has not moved, locations must be counted using a technique which groups closely-spaced GPS location samples together and considers them as corresponding to the same location.

In order to estimate the number of locations visited from GPS location data, location data was first processed to identify stationary points. A stationary point is defined as one in which a subject has travelled at a speed less than 1 km/h relative to the last recorded location point in time (assuming direct-line travel). This stationary location data was then processed using a clustering approach [108], to group closely-located readings into distinct locations visited by the subject. The specific clustering algorithm used is called DBSCAN [109], parameterized with an epsilon value of 150 meters and a minimum samples value of 2. The Number of Locations Visited feature is then simply the number of clusters produced when clustering a participant's stationary GPS location data:

$$\text{Number of Locations} = |clusters(\text{SGPS})| \tag{7.3}$$

where SGPS refers to a participant's stationary GPS location data.

Figure 7.1 provides an illustration of one participant's location data. The left-most plot shows all location data samples collected, while the right-most plot shows only stationary samples, color-coded by their cluster. In total, 13 location clusters were identified in this instance. Please note that the axes of the plots are intentionally unmarked in order to preserve participant privacy. Also note that many closely-located samples appear as a single point in the plots in order to produce a more easy-to-read figure.

This feature has also been adopted from the prior work by Saeb et al in [44], which referred to this feature as *Number of Clusters*. It must be noted, however, that Saeb's work used the K-means clustering algorithm, and not the DBSCAN clustering algorithm. DBSCAN was chosen as an alternative because K-means operates by computing the Euclidean distance between the data to be clustered, which does not accurately measure the distance between two pairs of locations when represented as latitude and longitude (latitude and longitude represent angles, and the distance between two points on the surface of a sphere is not equal to the Euclidean distance between the angles themselves). Instead, the DBSCAN algorithm was chosen and used along with a Haversine (also known as great-circle) distance metric [110], which measures the distance between GPS locations as expected.

Figure 7.1: A study participant's raw and clustered location data.

Location Entropy Feature

The *Location Entropy* feature is intended to act as a measure of the variability of times spent at the different locations visited by participants. This feature was developed by Saeb et al in [44], and draws upon the concept of entropy from information theory. The location entropy feature is computed as

$$\text{Location Entropy} = -\sum_{i} p_i \ln p_i \tag{7.4}$$

where p_i is the percentage of time spent by the participant at a location i.

This feature achieves a maximum value when an equal amount of time is spent at each location, and achieves lower values when time is split unevenly between locations. In this sense, an individual which spends the majority of their time at their home would have a relatively low Location Entropy, while an individual which spends their time split more evenly between many locations would have a higher Location Entropy.

Normalized Location Entropy Feature

The Location Entropy feature, as defined above, can achieve a maximum value of $\log N$, where N is the number of locations visited by the individual, and therefore the value of this feature is influenced by the number of locations visited by an individual.

In order to break this dependence, the *Normalized Location Entropy* is defined as

$$\text{Normalized Location Entropy} = \text{Location Entropy} / \log N \tag{7.5}$$

were N is the number of locations visited by the participant. This feature was developed by Saeb et al in [44].

Home Stay **Feature**

The *Home Stay* feature measures the fraction of timethat participants spend at home. It was developed by Saeb et al in [44], and is computed as the ratio of the number of location samples which fall in the participant's home location cluster to the total number of location samples collected.

$$\text{Home Stay} = \frac{|\text{CGPS}_{cluster=Home}|}{|\text{GPS}|} \tag{7.6}$$

where CGPS refers to the set of participant's *clustered* GPS location samples, and GPS refers to the participant's set of (unclustered) GPS location samples.

A participant's home location cluster is inferred to be the cluster which is most visited (i.e., has the largest number of location samples assigned to it) between the hours of 12 AM and 6 AM.

Circadian Movement **Feature**

The *Circadian Movement* feature measures the degree to which a participant's pattern of travel to different locations follows a repeating, 24-hour rhythm. This feature was developed by Saeb et al in [44]. Conceptually, an individual which visits the same locations at the same times every day would have high Circadian Movement. Mathematically, this is measured by using a form of spectral analysis called the Lombe-Scargle method [111], which is a technique that can be used to estimate the frequency spectrum of a signal. The Lombe-Scargle method is preferable to the more commonly-used Fourier analysis [112] in this setting since it easily handles missing data and data which is not sampled with perfect periodicity (i.e., different length of time between samples), both of which are true for the data collected in this work.

The Lombe-Scargle method was used to estimate the power spectral density (PSD) of the GPS location data, by first computing the PSD of the longitude and latitude data independently. Using each PSD, the power of the signals which fell between the frequency bins corresponding to a period of 24 ± 0.5 hours was measured:

$$P = \sum_{i=1.13 \times 10^{-5}\,\text{Hz}}^{1.18 \times 10^{-5}\,\text{Hz}} PSD(f_i)$$

where the frequencies of 1.13×10^{-5} Hz and 1.18×10^{-5} Hz correspond to periods of 24.5 Hours and 23.5 Hours, respectively.

Power was measured across these frequency bins in order to assess the spectral content of the signal near to having a frequency of once per day, with some leeway to account for small variability in daily schedules. The Circadian Movement feature was then computed as:

$$\text{Circadian Movement} = log(P_{lat} + P_{lon}) \tag{7.7}$$

where P_{lat} and P_{lon} refer to the power of the latitude and longitude signals contained within the frequency bins corresponding to the periods of 24 ± 0.5 hours, respectively.

Transition Time **Feature**

The *Transition Time* Feature is a measure of the amount of time which individuals spend moving between different locations. This feature was developed by Saeb et al in [44], and is computed as the ratio of the number location samples which were non-stationary to the total number of location samples:

$$\text{Transition Time} = \frac{|\text{GPS}| - |\text{SGPS}|}{|GPS|} \tag{7.8}$$

where |GPS| is the set of GPS location samples collected from a participant and SGPS is the subset of *stationary* GPS location samples collected from participant. A location sample is deemed stationary if the time derivative of the distance to the last recorded sample is less than 1km/h (assuming direct-line travel).

Total Distance **Feature**

The *Total Distance* feature is a measure of the distance travelled by an individual. This feature was developed by Saeb et al in [44], and is computed as the sum of distances, measured in kilometres, between contiguous location samples:

$$\text{Total Distance} = \sum_{i\ \epsilon\ \text{GPS}} distance(i, i+1) \tag{7.9}$$

where GPS is the set of GPS location samples collected from a participant, and i

is the i^{th} sample in the set.

Exits from the Home **Feature**

The *Exits from the Home* feature is a proxy measure of participants' activity and social engagement. The feature counts the number of times that a participant left their home. It is computed by counting the number of transitions over time from the home location to any other location cluster:

$$\text{Exits from the Home} = \sum_{i \ \epsilon \ \text{CGPS}} Exit(i, i+1) \tag{7.10}$$

where CGPS is the subset of clustered GPS location samples collected from a participant and $Exit(a, b)$ is a function that evaluates to 1 if a GPS location point a was assigned to the home cluster and GPS location point b was not:

$$Exit(a, b) = \begin{cases} 1 & \text{if } cluster(a) = Home \ \& \ cluster(B)! = Home \\ 0 & \text{otherwise} \end{cases}$$

This feature is similar to the Home Stay feature, which measures the amount of time spent at home, but was designed to provide some additional insight into avoidance behavior beyond what the Home Stay feature might capture. Consider, for example, two individuals (Individuals "A" and "B") which both spend 80% of their time at home. Individual A may choose to leave the home very infrequently, by performing all their necessary errands in one day, after which they are free to continue to self-isolate for a number of days by not leaving the home. Individual B, however, may be less averse to leaving the home and does not "batch" their errands in this manner, leaving the home more frequently to perform their errands whenever necessary. In this example, both individuals would have equivalent Home Stays, but different Exits from the Home.

Home Stay Entropy **Feature**

The *Home Stay Entropy* Feature is a measure of the variability in the length of times spent at home. To compute this feature, a histogram of the times spent at home is created, binning these durations into 336 bins, each one hour wide, in order to represent durations in the range from 1 hour to 336 hours (i.e., the total duration of the 14-day study). Home stay entropy is then computed as

$$\text{Home Stay Entropy} = \sum_{i=0}^{335} p_i \log p_i \tag{7.11}$$

where p_i is equal to the number of times spent at home with a duration in the range of $(i, i+1]$ hours divided by the total number of times the individual was at home.

This feature is minimized when all times at home are of an equal duration, and therefore higher values of this feature indicate more variance in individual home stay durations. Note that no normalization of this feature is necessary since we are computing the entropy of a distribution over *time*, in contrast with the Location Entropy feature, where entropy of a distribution over *locations* was being computed. All participants remain in the study for the same duration of time, but do not visit the same number of locations.

Summary of Measured Feature Values

Summary statistics describing the distribution of participants' feature values are provided in Table 7.4. Inspection of the Number of Locations feature reveals that, on average, participants visited 15.2 different locations over the duration of the study. The mean value of the Home Stay feature indicates that, on average, participants remained at home 70% of the time. Shifting attention to the Transition Time feature, participants spent, on average, 10% of the time moving between locations. The average Total Distance travelled by participants was 678 km. Finally, the average number of Exits from the Home was 15, indicating that participants left the home roughly once per day.

Feature	Mean (SD)	Min.	First Quartile	Second Quartile	Third Quartile	Max.
Location Variance	-6.3 (4.5)	-19.5	-8.3	-6.5	-4.2	5.9
Number of Locations	15.2 (9.9)	1.0	8.0	13.0	20.5	50.0
Location Entropy	0.7 (0.4)	0.0	0.5	0.7	1.0	2.0
Normalized Location Entropy	0.3 (0.2)	0.0	0.2	0.3	0.3	0.8
Home Stay	0.7 (0.2)	0.3	0.6	0.7	0.8	1.0
Circadian Movement	4.3 (2.0)	-2.7	3.3	4.6	5.8	6.8
Transition Time	0.1 (0.0)	0.0	0.0	0.1	0.1	0.3
Total Distance	677.8 (994.8)	5.6	153.0	332.7	888.0	6944.5
Exits from Home	15.0 (6.9)	1.0	11.0	14.0	18.5	34.0
Home Stay Entropy	2.0 (0.5)	0.0	1.9	2.1	2.4	2.7

Table 7.4: Summary statistics describing participants' location features (n=71).

7.2.3 Correlations with Mental Health Measures

To understand how well these features actually captured any symptoms of anxiety, depression, or overall poor mental health, correlations between all features and all

self-report measures were computed. Table 7.5 shows the results of this correlation analysis.

Feature	Correlations with Mental Health Measures: r (P)			
	LSAS	GAD-7	PHQ-8	SDS
Location Variance	-0.04 (0.728)	-0.05 (0.669)	-0.24 (0.042)	-0.30 (0.011)
Number of Locations	-0.24 (0.046)	-0.26 (0.026)	-0.30 (0.010)	-0.28 (0.017)
Location Entropy	-0.09 (0.471)	-0.05 (0.674)	-0.24 (0.048)	-0.20 (0.087)
Normalized Location Entropy	-0.05 (0.676)	0.05 (0.650)	-0.15 (0.222)	-0.09 (0.449)
Home Stay	0.12 (0.320)	0.00 (0.972)	0.19 (0.113)	0.22 (0.067)
Circadian Movement	0.09 (0.479)	0.09 (0.467)	0.04 (0.759)	0.08 (0.511)
Transition Time	-0.27 (0.024)	-0.30 (0.012)	-0.34 (0.004)	-0.33 (0.005)
Total Distance	0.06 (0.622)	0.04 (0.748)	-0.09 (0.441)	-0.22 (0.071)
Exits from Home	-0.25 (0.036)	-0.31 (0.009)	-0.29 (0.014)	-0.20 (0.087)
Home Stay Entropy	-0.25 (0.036)	-0.26 (0.031)	-0.31 (0.008)	-0.21 (0.083)

Table 7.5: Correlations between location-derived features and mental health measures (n=71).

The Location Variance feature was significantly correlated (at a 5% significance level) with PHQ-8 scores (r = -0.24, P = 0.04) and SDS scores (r = -0.30, P = 0.01). This negative correlation indicates that individuals with more variance in their locations had *lower* self-reported symptoms of depression and functional impairment, a result which is in line with our hypotheses. Similarly, the Number of Locations feature had significant correlations with the GAD-7 (r = -0.26, P = 0.03), the PHQ-8 (r = -0.30, P = 0.01), and the SDS (r = -0.28, P = 0.02). Transition Time had significant, negative correlations with the LSAS (r = -0.27, P = 0.02), GAD-7 (r = -0.30, P = 0.01), PHQ-8 (r = -0.34, P = 0.01), and the SDS (r = -0.33, P = 0.01). The Exits from the Home feature had significant negative correlations with the LSAS (r = -0.25, P = 0.04), GAD-7 (r = -0.31, P = 0.01), and PHQ-8 (r = -0.29, P = 0.01). Similarly, the Home Stay Entropy feature had significant negative correlations with the LSAS (r = -0.25, P = 0.04), GAD-7 (r = -0.26, P = 0.03), and PHQ-8 (r = -0.31, P = 0.01).

7.2.4 Discussion

Key Findings

The Number of Locations feature, when compared to the Location Variance feature, additionally has a significant correlation with the GAD-7 and a near-significant correlation with the LSAS. This may indicate that the Number of Locations feature not only captures physical activity, but some degree of social avoidance, or more accurately, a lack of social avoidance. While no distinction is made between locations, and we therefore do not know if individuals are visiting locations which are social or otherwise likely to trigger anxiety, it may be enough to know the number of locations

visited to gain some insight into the degree of avoidance behavior, or lack thereof.

The Home Stay feature, which is a measure of the time spent at home, failed to have significant correlations with mental health measures, while the Exits from Home feature did. As was mentioned in the motivating example behind the Exits from the Home feature, knowledge of how many times individuals left the home, and not just how much time they spend at home, may offer enhanced insight into mental state.

A general finding that is worthy of emphasis here is that four of the location-based features (Number of Locations, Transition Time, Exits from the Home, and Home Stay Entropy) had significant correlations with measures of anxiety in addition to significant correlations with measures of depression. The link between energy, movement, and activity and depression is quite clear and prominent, considering that it appears directly in the diagnostic criteria for depression. That location-based features have the ability to quantify these characteristics and therefore correlate with a measure of depression is perhaps not surprising. That these features additionally capture some degree of anxious symptomatology is more surprising, especially since the Number of Locations feature and the Transition Time feature were both originally designed in a study solely focused on depression.

Comparison with Prior Work

Considering that the first nine of eleven features described in this section were all adopted from the previous work in [44], it may be illuminating to compare and contrast how well these features correlate with participants' self-reported measures of depression in the two studies. We are also able to compare to Saeb's own replication study [55], which sought to replicate the original results on a new data set.

Table 7.6 presents the correlations (Pearson r) between location features and self-reported symptoms of depression. Note that both Saeb studies measured depression using the 9-item PHQ-9 questionnaire, while this work used the reduced 8-item PHQ-8 questionnaire (which lacks the final question assessing suicidality and self-harm). Also note that the three studies have different sample sizes: 18, 48, and 71 participants, respectively. P-values are absent from the results of the Saeb 2016 study; 95% confidence intervals were reported instead.

A general trend when comparing these results is that, in all but two cases, the correlations all share the same directionality. There are two outliers to this trend. The first is that this work measures a insignificant, positive correlation between Circadian Movement and PHQ-8 scores, while both prior works found a strong negative correlation with PHQ-9 scores. The second is the Transition Time feature in the original 2015 study, which had a positive correlation with PHQ-9 scores, while a negative

Feature	Correlation with PHQ-9/PHQ-8: r (P)		
	Saeb 2015 (n = 18)	Saeb 2016 (n = 48)	This Work (n = 71)
Location Variance	-0.58 (0.01)	-0.43 ± 0.007	-0.24 (0.04)
Number of Locations	-0.09 (0.73)	-0.44 ± 0.004	-0.30 (0.01)
Location Entropy	-0.42 (0.08)	-0.46 ± 0.005	-0.24 (0.05)
Normalized Location Entropy	-0.58 (0.01)	-0.44 ± 0.005	-0.15 (0.22)
Home Stay	0.49 (0.04)	0.43 ± 0.005	0.19 (0.11)
Circadian Movement	-0.63 (0.01)	-0.48 ± 0.006	0.04 (0.76)
Transition Time	0.21 (0.40)	-0.32 ± 0.005	-0.34 (0.01)
Total Distance	-0.08 (0.77)	-0.18 ± 0.006	-0.09 (0.44)

Table 7.6: Comparison of correlations between location-derived features and self-reported symptoms of depression reported in this work and two prior works.

correlation was measured in both the 2016 study and this work.

A second general trend is that the strength of the correlations reported in this work are, in almost all cases, weaker. The Location Entropy, Normalized Location Entropy, Home Stay, and Circadian Movement features all had strong ($|r| > 0.4$), significant correlations with PHQ-9 scores in both Saeb studies, but failed to be strong enough to achieve significance (at a 5% significance level) in this work.

One key factor which may account for the weaker correlations is related to the sample of participants in each of the studies. While it is unclear where participants of the Saeb 2015 study lived (i.e., rural or urban areas), participants in the 2016 study were all students at the same University. Contrast with this work, where participants were recruited from all across Canada, from rural and urban areas. Furthermore, the sample of participants in this work includes both students (35%) and non-students (65%). The heterogeneity of this sample, especially with respect to area of residence, is especially relevant for location-based analysis, since patterns of travel and activity can vary significantly depending on the nature of the area in which one lives. The relationship between some of these location-based features and depression may be moderated by the nature of the location in which individuals live.

A final factor which may account for the observed differences in effect sizes between our work and both Saeb studies is the sampling period at which GPS location data was collected. The Saeb 2015 study asampled GPS location data with a period of 5 minutes, while the Saeb 2016 study consisted of data which was sampled with a period of 10 minutes. In this work, the 75th percentile of average GPS location sampling periods was 13.0 minutes. The reduced sampling rate in this subset of participants could potentially account for some weakening of associations. Note that the data in both Saeb studies was collected prior to 2015, at which point Android 6 was released, which introduced Doze mode and the inability to perform precisely-timed background

data collection (see Section 6.4.1 for a discussion of the challenges in post-Android 6 passive data collection).

7.3 Conclusion

This chapter has described the design of features derived from activity recognition and GPS location data. The novel *Activity Rate* feature, extracted from activity recognition data, was shown to be correlated with symptoms of social anxiety (r = -0.24, P = -.03) and depression (r = -0.26, P = 0.02). A number of location-based features described in prior work were investigated, replicating previous results which found correlations between these features and symptoms of depression and, additionally, it was shown that some of these features are also correlated with symptoms of anxiety. Two novel location-based features —Exits from Home and Home Stay Entropy —were also designed and shown to correlate with symptoms of anxiety and depression. The next chapter describes the design of features derived from battery charge, light sensor, and screen activity data.

Chapter 8

Battery, Light, and Screen-Based Features

This chapter continues the analysis of the smartphone-collected data, in which features are designed to capture mental health signals from each of the data streams collected. This chapter focuses upon the design and extraction of features derived from participants' battery charge, light sensor, and screen activity data. The features presented in this chapter are a mix of novel features and features which replicate prior work; replicated features will cite the prior work in which they were first designed and presented.

8.1 Battery Charge Data

8.1.1 Data Overview

The battery charge level of participants' smartphones was nominally sampled by the smartphone application every 5 minutes, but the effective sampling period achieved by the application was, on aggregate, longer than this (the 75th percentile of average sampling periods of battery charge data was 7.5 minutes) As with all data streams, a threshold on the minimum number of samples per participant for inclusion in analysis was set in order to handle missing data. This threshold is identical to the threshold used in the analysis of the Activity Recognition data; participants are required to have at least half the nominal number of samples expected (2,016 samples). Applying this threshold results in a sample size of 84 participants for analysis.

8.1.2 Analysis and Feature Design

The trend of participants' battery charge over time contains some characteristics of potential interest. Firstly, it is clear when a device's battery is charging or depleting, by noting whether the charge follows an increasing or decreasing trajectory, respectively. Furthermore, the number of times that a participant charges their device is also evident by noting how many times the charge dips to a local minimum and then begins to increase again. One can also gain some sense of the average charge level of the device. Figure 8.1 provides a visualization of one participant's battery charge data where all these characteristics are visible.

All of these pieces of information may potentially give some insight into aspects of an individual's mental health. The key hypothesis that drove the creation of features derived from battery charge data was that anxious individuals may be less willing to let their device charge dip too low. Anxious individuals may have a tendency to avoid the risk of being without a usable smartphone. This seems plausible when viewed through the lens of the psychological construct known as *intolerance of uncertainty*. Intolerance of uncertainty represents an individual's negative emotional response to situations where outcomes are uncertain [113], and intolerance of uncertainty has been suggested as a risk factor for the development of anxiety disorders [114].

It is possible that anxious individuals could view the situation of having a depleted smartphone battery as highly uncertain, since they could miss important calls or messages, or be unable to get help in the case of an emergency. In this sense, anxious individuals may be averse to letting their smartphone's battery charge dip too low and would be more vigilant in charging it. Three features were extracted from battery charge data which all attempt to quantify this behavior.

Average Charge **Feature**

The *Average Charge* feature is intended to give a rough, aggregate measure of the charge level of a participant's smartphone over the duration of the study. A higher Average Charge may indicate that a participant is more concerned about low battery charge. Average Charge is computed simply as the average of all battery charge samples:

$$\text{Average Charge} = mean(\text{BC}) \tag{8.1}$$

where BC is the set of battery charge samples collected from a participant.

Number of Charges Feature

The *Number of Charges* feature measures how many times a participant charged their device over the duration of the study. More frequent charging of a device might again indicate that an individual is more concerned about low battery charge. This feature is computed by counting the number of local minima contained within a participant's battery charge data over time:

$$\text{Number of Charges} = |local\ minima(\text{BC})| \tag{8.2}$$

where BC is the set of battery charge samples collected from a participant.

Average Minimum Charge Feature

The *Average Minimum Charge* feature measures how low, on average, the battery charge level of the device dropped each time the participant began to charge the device. An individual with a high Average Minimum Charge may be more worried about low battery charge, and charge their phone earlier than other, less worried individuals, do. This feature is computed as the average battery charge level at the local minima identified when computing the Number of Charges feature.

$$\text{Average Minimum Charge} = mean(local\ minima(\text{BC})) \tag{8.3}$$

where BC is the set of battery charge samples collected from a participant.

Summary of Measured Feature Values

Figure 8.1 provides a visualization of one participant's battery charge data, with the corresponding values of the Average Charge, Number of Charges, and Average Minimum Charge features also annotated.

Summary statistics describing the distribution of participants' feature values are provided in Table 8.1. The mean Average Charge was 68%. The mean Number of Charges was 32, indicating that participants charged their device roughly twice a day over the 14-day study. Participants waited until their devices hit a low charge level of 59%, on average, before re-charging their devices.

8.1.3 Correlations with Mental Health Measures

To understand how well these features actually captured any symptoms of anxiety, depression, or overall poor mental health, correlations between all features and all self-report measures were computed. Table 8.2 shows the results of this correlation

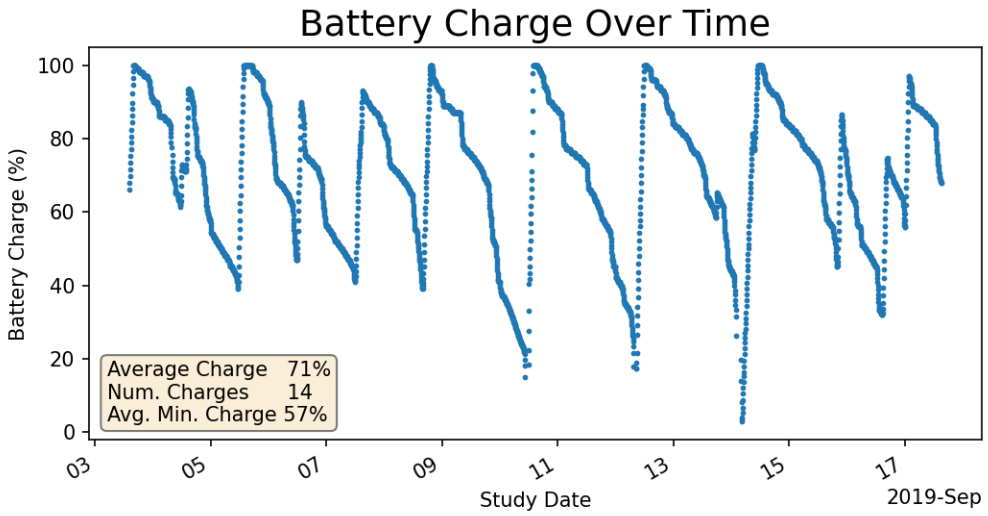


Figure 8.1: A specific study participant's battery charge data annotated with corresponding feature values.

Feature	Mean (SD)	Min.	First Quartile	Second Quartile	Third Quartile	Max.
Average Charge	68 (12)	35	61	69	77	99
Number of Charges	32 (18)	2	19	29	39	102
Average Minimum Charge	59 (18)	10	49	59	73	91

Table 8.1: Summary statistics describing participants' battery charge features (n = 84).

analysis. The Average Charge feature was essentially uncorrelated with any mental health measure. The Number of Charges feature also had very low correlations (all correlations less than or equal to 0.10). The Average Minimum Charge feature had weak, positive correlations with all mental health measures. This could indicate that individuals with worse mental health cannot tolerate low charge levels before choosing to charge their phones, but the lack of any significant correlations (at a 0.05 significance level) makes this an uncertain claim.

Feature	Correlations with Mental Health Measures: r (P)			
	LSAS	GAD-7	PHQ-8	SDS
Average Charge	0.00 (0.98)	-0.06 (0.61)	0.04 (0.75)	-0.01 (0.94)
Number of Charges	0.07 (0.50)	0.10 (0.35)	0.03 (0.80)	0.06 (0.58)
Average Minimum Charge	0.11 (0.32)	0.14 (0.19)	0.12 (0.29)	0.19 (0.08)

Table 8.2: Correlations between battery-derived features and mental health measures (n = 84).

8.1.4 Discussion

One key limitation, and possible confounding effect, could account for why these features failed to have any significant correlations with mental health measures. Namely, that different devices have different rates at which they lose their battery charge. For example, a new device with a large, healthy battery can offer much more battery life at a low charge than an older device with a smaller or worn-out battery. If the hypothesis is that individuals cannot tolerate the risk of having their device lose charge, a more appropriate battery feature would estimate the remaining device usage time (given the charge level) at the time of charging, instead of the charge level itself.

8.2 Light Sensor Data

8.2.1 Data Overview

The illuminance of participants' environments, as measured by their smartphone's light sensor, was nominally sampled by the smartphone application every 10 minutes, but the effective sampling period achieved by the application was, on aggregate, longer than this (the 75th percentile of average sampling periods of light sensor data was 11.1 minutes). As with all data streams, a threshold on the minimum number of samples per participant for inclusion in analysis was set in order to handle missing data. This threshold is similar to the threshold used in the analysis of the Activity Recognition data; participants are required to have at least half the nominal number of samples expected (1,008 samples). Applying this threshold results in a sample size of 83 participants for analysis.

8.2.2 Analysis and Feature Design

The hypothesis which drove the creation of features derived from light sensor data was that depressed individuals would find themselves in environments which had distincly different lighting conditions than healthy individuals. This is derived from two key characteristics of depression: hypersomnia and troubled sleep [17]. Hypersomnia might be accompanied by more time spent in dark environments, assuming that individuals sleep in the dark. Troubled sleep might be accompanied by a slightly less dark environment at night, since individuals which wake up in the middle of the night might turn on the lights briefly to read, check their phone, or grab a drink of water, for example.

Average Illuminance Feature

The *Average Illuminance* feature is intended to give a rough, aggregate measure of the brightness of a participant's environment over the duration of the study. Average Illuminance is computed simply as the average of all light sensor samples:

$$\text{Average Illuminance} = mean(\text{LS}) \tag{8.4}$$

where LS is the set of light sensor samples collected from a participant.

Weeknight Illuminance Feature

In order to build a proxy measure of sleep disturbance, the brightness of the environment during common sleep times is directly assessed here. The *Weeknight Illuminance* feature measures the average illuminance of a participant's environment, during common hours of sleep on weeknights. This feature is restricted to weeknights only since participants may have different sleep schedules on weekend evenings (especially considering how many young participants exist within the sample). Weeknight Illuminance is computed simply as the average of the illuminance data points captured on weeknights between the hours of 12:00 AM and 6:00 AM:

$$\text{Weeknight Illuminance} = mean(\text{LS}^{weeknight}) \tag{8.5}$$

where $\text{LS}^{weeknight}$ is the subset of light sensor samples collected from a participant on weeknights between the hours of 12:00 AM and 6:00 AM.

Time in Darkness Feature

The *Time in Darkness* feature measures the proportion of time that a participant's device is in a dark environment. This feature is computed as the percentage of light sensor data samples that measured an illuminance of less than 5 lux, a value which corresponds to a dark environment.

$$\text{Time in Darkness} = \frac{|\text{LS}_{ill.<5\ lux}|}{|\text{LS}|} \times 100\% \tag{8.6}$$

where LS is the set of light sensor samples collected from a participant.

Summary of Measured Feature Values

Figure 8.2 provides a visualization of one participant's light sensor data, with the corresponding values of the Average Illuminance, Average Weeknight Illuminance, and Time in Darkness features also annotated.

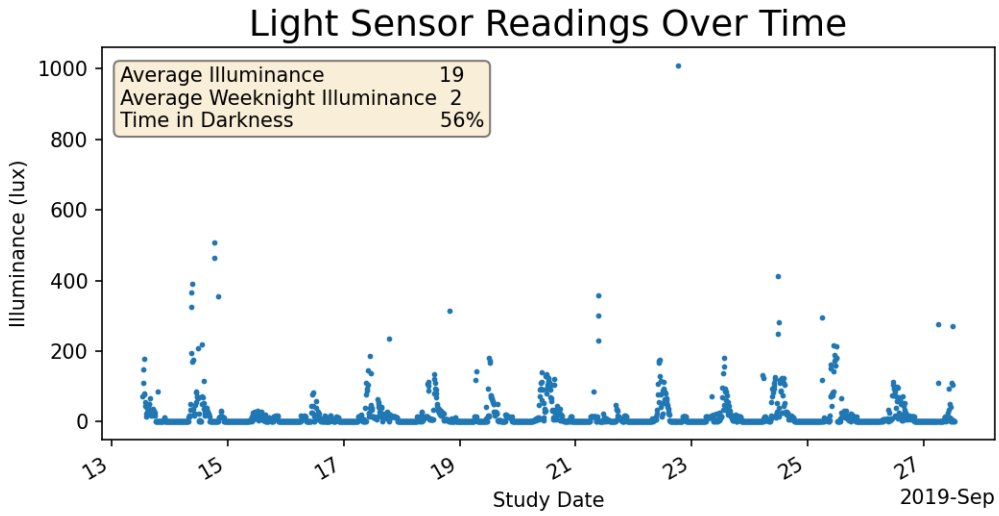


Figure 8.2: A specific study participant's light sensor data annotated with corresponding feature values.

Summary statistics describing the distribution of participants' feature values are provided in Table 8.3. The mean Average Illuminance was 74.4 lux. The mean Average Weeknight Illuminance was much lower, as would be expected, at 6.5 lux. The mean value of the Time in Darkness feature was 62.5%, which indicates that, on average, participants' devices measured a dark environment the majority of the time.

Feature	Mean (SD)	Min.	First Quartile	Second Quartile	Third Quartile	Max.
Average Illuminance	74.4 (86.7)	2.9	27.5	49.5	85.0	625.8
Average Weeknight Illuminance	6.5 (16.4)	0.0	0.1	1.2	5.4	95.8
Time in Darkness	62.5 (14.1)	30.5	51.2	62.1	73.8	96.8

Table 8.3: Summary statistics describing participants' light features (n = 83).

8.2.3 Correlations with Mental Health Measures

To understand how well these features actually captured any symptoms of anxiety, depression, or overall poor mental health, correlations between all features and all self-report measures were computed. Table 8.4 shows the results of this correlation analysis. The Average Illuminance feature had weak, negative correlations with all mental health measures. The Average Weeknight Illuminance had a weak, positive correlation with GAD-7 scores (r=0.16, P=0.15). The Time in Darkness feature was essentially uncorrelated with all mental health measures ($r < 0.08$ in all cases). No measured correlations were statistically significant at a $\alpha = 0.05$ significance level.

Feature	Correlations with Mental Health Measures: r (P)			
	LSAS	GAD-7	PHQ-8	SDS
Average Illuminance	-0.10 (0.38)	-0.10 (0.38)	-0.14 (0.22)	-0.12 (0.28)
Average Weeknight Illuminance	-0.06 (0.57)	0.16 (0.15)	0.02 (0.88)	0.04 (0.74)
Time in Darkness	-0.02 (0.87)	0.05 (0.65)	0.07 (0.53)	0.05 (0.68)

Table 8.4: Correlations between light-derived features and mental health measures (n = 83).

8.2.4 Discussion

One key limitation may account for why these features failed to have any significant correlations with mental health measures. Namely, that the illuminance readings produced by smartphone light sensors do not always capture the lighting environment that individuals find themselves in. Consider, for example, a device which has its light sensor obscured by being in a pocket, in a bag, or by being placed face down on a desk or table (light sensors are on the face of the device). Such a device will necessarily produce illuminance values corresponding to a dark environment, while the environment that the participant is in may be well-lit. Given this limitation, conclusions regarding the associations between dark environments and mental health cannot be drawn from this data alone.

8.3 Screen Activity Data

8.3.1 Data Overview

The state of participants' smartphones screens (i.e., on or off) was not sampled periodically, but was instead recorded on an event-driven basis. Each time the smartphone entered an interactive state, due to turning on the screen by pressing the power button, or receiving a notification, the study smartphone application recorded this as a "screen on" event. Similarly, each time a smartphone entered an un-interactive state, by turning the screen off or by automatically entering a sleep state, the application recording this as a "screen off" event.

As with all data streams, a threshold on the minimum number of samples per participant for inclusion in analysis was set in order to handle missing data. As opposed to the periodically-sampled data streams, the event-driven nature of this data collection makes identifying missing data difficult. Long periods of time with no recorded screen state changes (e.g., a whole day or more) could correspond either to a participant simply not interacting with their device, or to a loss of data (i.e., the particular module of the study application which listens for this data was temporarily killed by the operating system).

A heuristic was used to identify and remove participants from analysis who were likely to be missing screen activity data. Missing screen activity data would most likely occur along with missing data in other streams, since suspension of one data-collecting module of the study application will often be accompanied by the suspension of the others. Therefore, a threshold was applied to the number of audio recordings captured. Participants were required to have half the nominal number of audio recordings expected (2,016 recordings). Applying this threshold results in a sample size of 84 participants for analysis. Note this thresholding could have been performed on the basis of data yields in other streams with little difference to the number of participants excluded, since thesholding on Activity Recognition, Battery, Bigrams, or Light would all yield anywhere from 83-86 participants (the one outlier to this trend would be Location data, the thresholding of which yielded only 71 participants).

8.3.2 Analysis and Feature Design

To aid analysis, it is useful to visualize the screen activity data. One way to visualize this data is to build a heatmap of a participant's screen activity over time. Figure 8.3 shows one participant's screen activity over time, where the percentage of the time that the screen was active is plotted for each day and hour of the study. The intensity of the color at each $(day, hour)$ grid location corresponds to the percentage of time that the screen was on during that hour and on that day of the study.

Inspection of this participants' screen activity over time reveals some characteristics of potential interest. Most prominently, it is clear that the smartphone is not being used between the hours of midnight and 7 AM. At all other times, however, the device sees consistent activity, hovering at what appears to be about 40% or so.

To direct the design of features which may offer insight into individuals' levels of anxiety and depression, let us draw upon some knowledge from the mental health domain. Firstly, there is a known link between sleep and mental health, as was discussed in the section on light-derived features (Section 8.2.2). Given that an individual which is using their phone cannot be asleep, screen activity data was used to generate features which are proxy measures of sleep and sleep quality. Secondly, there is evidence of an association between general smartphone use and poor mental health. A study by Thomee et al. revealed that high mobile phone use was associated with symptoms of depression and also with sleep disturbances [115]. Compulsive smartphone usage was also found to be associated with stress and symptoms of social interaction anxiety in [116]. The features which follow were designed to capture and quantify smartphone usage and sleep patterns.

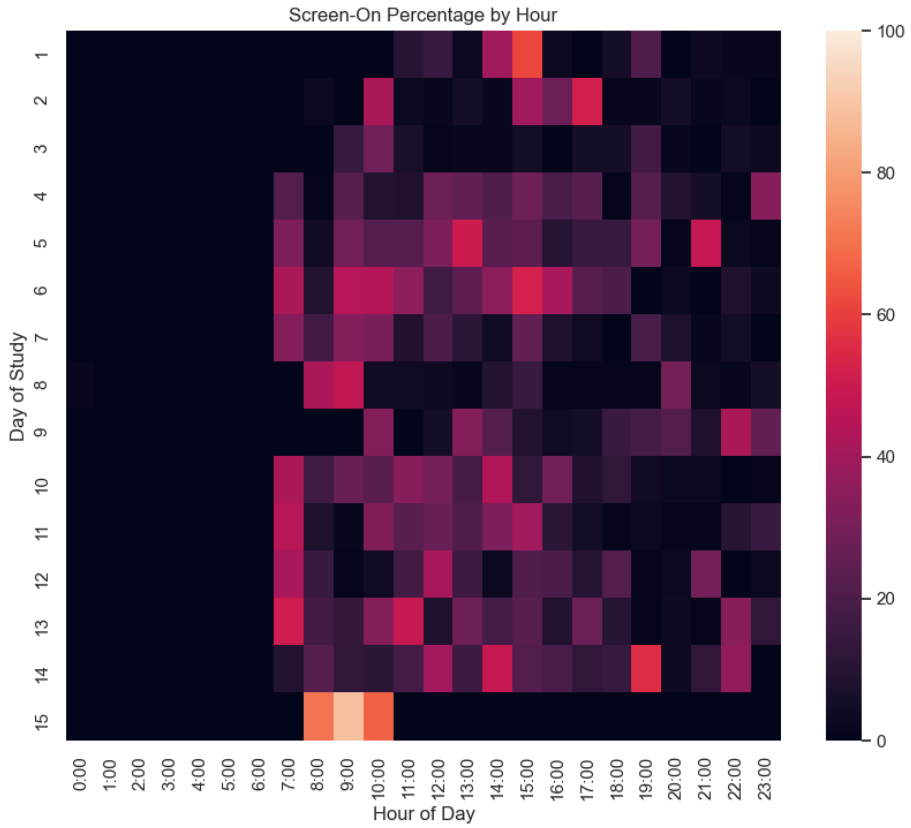


Figure 8.3: A specific study participant's screen activity over time in the study.

Screen Turn-Ons **Feature**

The *Screen Turn-Ons* feature is intended to be one measure of smartphone usage. It is simply a count of the number of times that the smartphone screen turned on:

$$\text{Screen Turn-Ons} = |\text{SA}_{state=ON}| \tag{8.7}$$

where SA is the set of screen activity samples collected from a participant. This feature is inspired by the Phone Usage Frequency feature designed by Saeb et al. in [44].

Screen Usage **Feature**

The *Screen Usage* feature is another measure of smartphone usage. It is computed as the fraction of time that the screen was on:

$$\text{Screen Usage} = \frac{t_{on}}{t_{on} + t_{off}} \tag{8.8}$$

where t_{on} and t_{off} are equal to the duration of time during which a participant's screen was On and Off, respectively. This feature is inspired by the Phone Usage Duration feature designed by Saeb et al. in [44].

Screen Usage Entropy **Feature**

The *Screen Usage Entropy* feature is a measure in the variability of smartphone usage across time. It is computed as the information-theoretical entropy of the distribution of screen usage when computing screen usage per hour (i.e., the fraction of time that the screen was on for each hour of the study).

$$\text{Screen Usage Entropy} = \sum_{i=1}^{14\times 24} p_i \log p_i \tag{8.9}$$

where p_i is the equal to the fraction of time that the screen was on during hour i of the study, normalized by the sum total of all hourly screen usage fractions. Figure 8.3 is a visualization of the (unnormalized) hourly screen usage values for one study participant.

An individual with consistent screen usage across all hours would have a high Screen Usage Entropy value, while an individual which only exhibits screen usage at select hours would have a low Screen Usage Entropy value. It is for this reason that the Screen Usage Entropy can be thought of as a measure of the variability of smartphone usage across time.

Nighttime Screen Usage **Feature**

The *Nighttime Screen Usage* feature is intended to be one proxy measure of sleep quality, by measuring the amount of screen activity during nighttime hours. It is computed as the fraction of time that the screen was on between the hours of midnight and 6 AM:

$$\text{Screen Usage} = \frac{t_{on}^{night}}{t_{on}^{night} + t_{off}^{night}} \tag{8.10}$$

where t_{on}^{night} and t_{off}^{night} are equal to the duration of time during which a participant's screen was On and Off, respectively, between the hours of midnight and 6 AM.

Weeknight Screen Usage **Feature**

To account for the fact that some individuals may stay up later and be more active during Friday and Saturday evenings, the *Weeknight Screen Usage* Feature is com-

puted as the fraction of time that the screen was on between the hours of midnight and 6AM on weeknights only:

$$\text{Weeknight Screen Usage} = \frac{t_{on}^{weeknight}}{t_{on}^{weeknight} + t_{off}^{weeknight}} \tag{8.11}$$

where $t_{on}^{weeknight}$ and $t_{off}^{weeknight}$ are equal to the duration of time during which a participant's screen was On and Off, respectively, between the hours of midnight and 6 AM on weeknights only.

Summary of Measured Feature Values

Summary statistics describing the distribution of participants' feature values are provided in Table 8.5. Participants' smartphones had their screen turn on an average of 1224 times. Note that turn-on events need not be initiated by the user, and can be triggered by receiving a notification, for example. Inspection of the Screen Usage features reveals that participants' smartphones were, on average, active 23% of the time, and this activity rate drops to 13%, on average, when only considering nighttime and weeknights. There was little difference in the distributions of the Nighttime Screen Usage and Weeknight Screen Usage features, which indicates that excluding Friday and Saturday evenings from the analysis, as was done in the case of Weeknight Screen Usage, was unnecessary.

Feature	Mean (SD)	Min.	First Quartile	Second Quartile	Third Quartile	Max.
Screen Turn-Ons	1224 (734)	128	772	1043	1600	3274
Screen Usage	0.23 (0.13)	0.02	0.12	0.21	0.29	0.60
Screen Usage Entropy	4.95 (0.44)	2.60	4.88	5.05	5.20	5.49
Nighttime Screen Usage	0.13 (0.15)	0	0.01	0.05	0.20	0.62
Weeknight Screen Usage	0.13 (0.16)	0	0.01	0.05	0.20	0.64

Table 8.5: Summary statistics describing participants' screen activity-based features (n = 84).

8.3.3 Correlations with Mental Health Measures

To understand how well these features actually captured any symptoms of anxiety, depression, or overall poor mental health, correlations between all features and all self-report measures were computed. Table 8.6 shows the results of this correlation analysis.

The Screen Turn-Ons and Screen Entropy features are essentially uncorrelated with any of the four self-report measures of mental health. Screen Usage, Nighttime Screen Usage, and Weeknight Screen Usage, were all *positively* correlated with all four

Feature	Correlations with Mental Health Measures: r (P)			
	LSAS	GAD-7	PHQ-8	SDS
Screen Turn-Ons	0.05 (0.65)	0.11 (0.33)	0.02 (0.83)	-0.03 (0.77)
Screen Usage	0.16 (0.15)	0.11 (0.31)	0.15 (0.17)	0.27 (0.01)
Screen Usage Entropy	0.11 (0.33)	0.01 (0.90)	-0.01 (0.96)	0.05 (0.68)
Nighttime Screen Usage	0.18 (0.10)	0.10 (0.38)	0.22 (0.04)	0.31 (0.01)
Weeknight Screen Usage	0.20 (0.07)	0.10 (0.37)	0.24 (0.03)	0.30 (0.01)

Table 8.6: Correlations between screen activity-derived features and mental health measures (n = 84).

mental health measures. This indicates that more smartphone usage was associated with worse mental health. Screen Usage was most strongly correlated with general impairment as measured by the SDS (r = 0.27, P = 0.01). The Nighttime Screen Usage feature achieved slightly stronger correlations with the LSAS, PHQ-8, and SDS than the general Screen Usage feature. The Nighttime and Weeknight Screen Usage features had near-identical correlations with mental health measures, as would be expected given that the features themselves are so closely related.

8.3.4 Discussion

The positive correlations between the Screen Usage Feature and all mental health measures coincide with previous findings that drew a connection between smartphone usage and poor mental health [115], [116]. Features which are comparable to the Screen Turn-Ons and Screen Usage features have both appeared in the prior work by Saeb [44]. In that study, a feature equivalent to Screen Turn-Ons was much more strong correlated with PHQ-9 scores (r = 0.52, P = 0.015). The same study also measured a much stronger correlation between smartphone usage and PHQ-9 scores (r = 0.54, P=0.011). To speculate on these differences, two distinctions between the our studies can be made. Firstly, the Saeb study had a very small sample size (21, as opposed to 84). Secondly, all participants in that study had PHQ-9 scores below 15. Of the 84 participants included in this analysis, 15 participants (18%) had a PHQ-8 score equal to or greater than 15. It may be that the relationship between smartphone usage and depressive symptoms is stronger for individuals with mild depressive symptoms. It may also be that the stronger correlations observed in [44] are due to chance, attributable to the small sample size.

8.4 Conclusion

This chapter has described the design of features derived from battery charge, light sensor, and screen activity data. All features derived from battery charge and light sensor data are novel designs, while the Screen Usage Entropy, Nighttime Screen Usage, and Weeknight Screen Usage features are also novel.

Battery charge and light-based features failed to show significant ($\alpha = 0.05$) correlations with measures of social anxiety, generalized anxiety, depression, or functional impairment. Screen activity-based features which quantified smartphone usage showed moderate correlations with symptoms of depression and functional impairment. The next chapter describes the design of features derived from audio recordings of participants' environments.

Chapter 9

Audio-Based Features

This chapter concludes the analysis of the smartphone-collected data, in which features were designed to capture mental health signals from each of the data streams collected. This chapter focuses upon the the design and extraction of features derived from participants' audio recordings, which yielded three constituent streams of data: audio volume, bigram, and voice activity data. The features presented in this chapter are a mix of novel features and features which replicate prior work; replicated features will cite the prior work in which they were first designed and presented.

9.1 Audio Volume Data

9.1.1 Data Overview

Audio volume is one of three streams of data extracted from audio recordings of participants' environments. This data is a measure of the average volume of the 15-second audio recordings of participants' environments. This data was sampled with a nominal sampling period of 5 minutes, but the effective sampling period achieved by the application was, on aggregate, longer than this (the 75th percentile of average sampling periods of audio volume data was 6.4 minutes)

As with all data streams, a threshold on the minimum number of samples per participant for inclusion in analysis was set in order to handle missing data. This threshold is identical to the threshold used in the analysis of the Activity Recognition data; participants are required to have at least half the nominal number of samples expected (2,016 samples). Applying this threshold results in a sample size of 84 participants for analysis.

Preprocessing of the volume time series was performed before feature extraction to account for missing data and to perform normalization. The audio volume time series

of each participant were resampled to a period of 5 min, and missing samples were imputed using linear interpolation between the two nearest samples. After resampling and interpolation, volume samples were clipped at a ceiling and floor of 3 standard deviations from the mean to remove outliers (using the mean and standard deviation of each individual participant's audio volume data set). Finally, the volume time series were scaled linearly to ensure that all volume measurements were within the range $[0, 1]$.

9.1.2 Analysis and Feature Design

To begin, it is useful to view a sample illustration of this audio volume data. Figure 9.1 shows a visualization of 7 days of one participant's audio volume data.

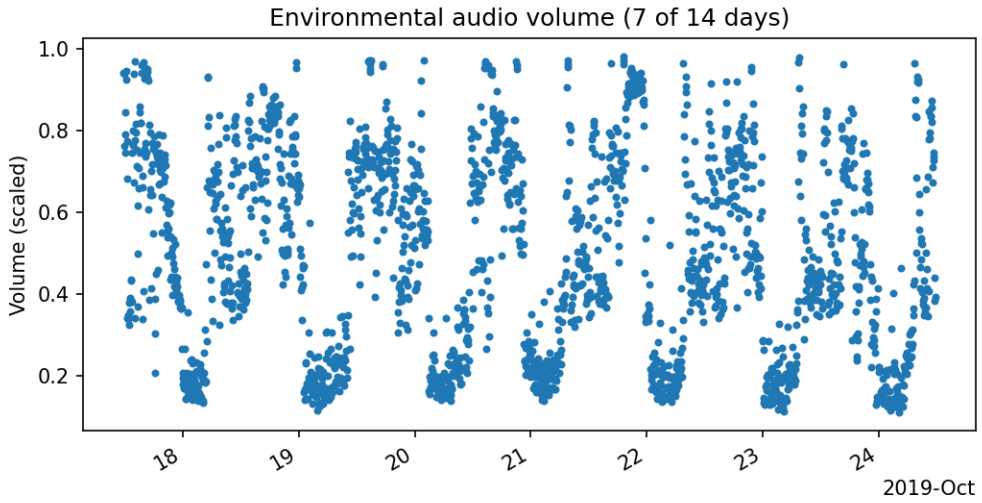


Figure 9.1: A specific study participant's audio volume data over time (7 of 14 days).

Visualizations of the volume time series show distinct periods of activity (characterized by large spikes in volume) and inactivity (characterized by quieter volume with less variance). These periods coincide roughly with daytime and nighttime, respectively. Two observations can be made about characteristics of this data which will be used to direct feature design. Firstly, this pattern of activity and inactivity is periodic and repeats daily. Secondly, the periods of apparent inactivity that are visible in the volume time series appear to coincide with the times that most people sleep. In an attempt to design features that are associated with and predictive of depression and anxiety, the mental health literature was surveyed for empirical or theoretical relationships between these characteristics and symptoms of anxiety and depression. If there is a known or suspected link between these characteristics and

depression or anxiety, then any features which quantify them may be useful for the purpose of predicting those disorders.

Focusing upon the periodic nature of the audio volume data, a link between this characteristic and participants' mental health might be found in the concept of social rhythms [117]. Social rhythm refers to the regularity in the pattern of social interactions and general activities. In essence, the ability to build and maintain a schedule is positive for mental health, and lower rhythmicity of daily life behaviors is associated with anxiety [118] and depression [119]. The rhythmicity of the apparent pattern of activity and inactivity in the audio volume data may act as a rough proxy measure of social rhythm, and may therefore be predictive of anxiety and depression.

The second observed quality of the audio volume data is that the periods of time with lower and less chaotic audio volume correspond roughly to sleep time. The quality of the audio during nighttime might give some indication of sleep quality, or of disturbances to sleep. Assessing sleep is a key area of focus for feature design since there are known links between sleep and mental health. The prevalence of mood disorders has been shown to be much higher in populations with chronic sleep problems [120], and insomnia may be a state marker of anxiety disorders [121]. It follows, therefore, that features which infer the quality of subjects' sleep may be associated with symptoms of anxiety and depression.

With an understanding of how both regularity in activity and sleep quality affect mental health, features were designed to quantify these phenomena.

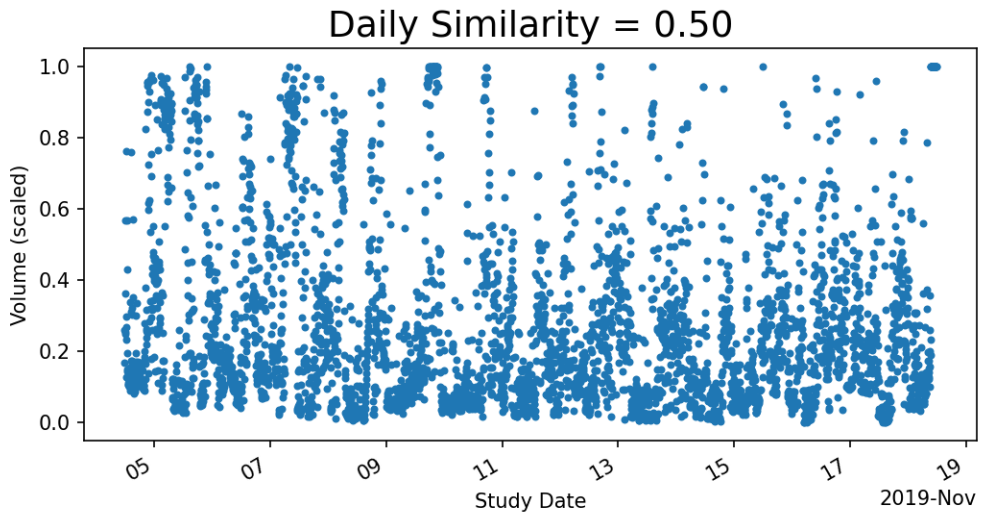
Daily Similarity **Feature**

The *Daily Similarity* feature was designed to infer the consistency of the sequence of participants' daily activities, by taking the peaks and troughs of the audio volume time series to represent periods of activity and inactivity, respectively. To quantify this consistency or regularity, the autocorrelation function was computed for each subject's volume time series. This is a signal processing technique that computes the correlation between a signal and a time-delayed copy of itself [122]. The autocorrelation function of a signal is a time-dependent Pearson correlation coefficient of the signal and its copy for varying degrees of time lag between the two. As we are interested in quantifying similarity across *days*, the value of the autocorrelation function at a time lag of 24 hours is most relevant. The Daily Similarity feature is, therefore, computed as the value of the autocorrelation function of the volume time series evaluated at a lag time of 24 hours:

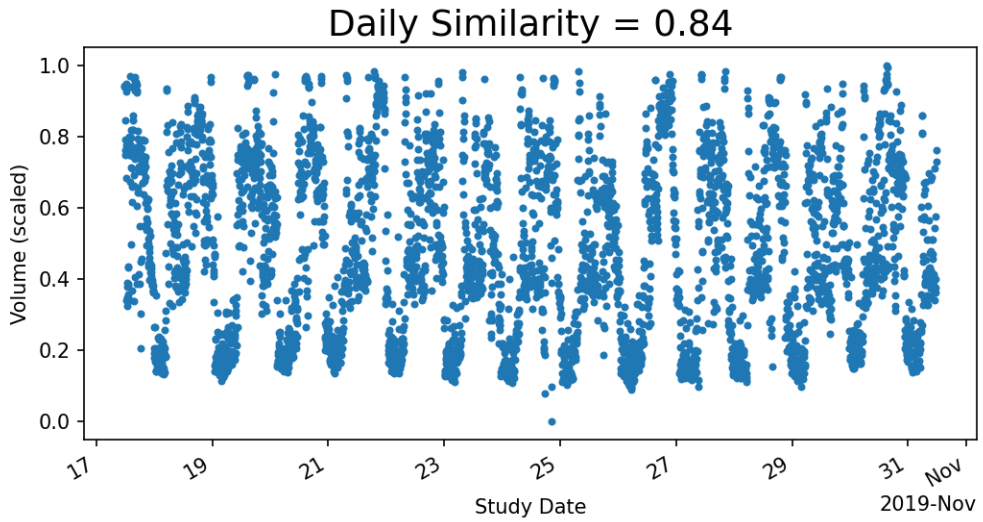
$$\text{Daily Similarity} = autocorr(\text{AV}, t_{24h}) \tag{9.1}$$

where AV represents the set of audio volume samples belonging to a participant.

Figure 9.2 provides a visualization of the audio volume belonging to two participants with contrasting Daily Similarity. The less regular data, shown in Figure 9.2a, yields a lower Daily Similarity than the more regular data, shown in Figure 9.2b.



(a) Audio Volume data with low corresponding Daily Similarity



(b) Audio Volume data with high corresponding Daily Similarity.

Figure 9.2: Audio volume data belonging to two participants with contrasting Daily Similarity.

Sleep Disturbance Feature

The *Sleep Disturbance* feature was designed to be a measure of participants' sleep quality. To quantify sleep quality, the volume time series was examined, with periods of quiet noted to be characterized by low variance in volume. Since noisy environments are more difficult to sleep in for most people, and a noisy environment might also indicate that a participant is awake and active, the noisiness of the environment during common hours of sleep can be considered a proxy measure of sleep quality.

To quantify noisiness of the environment, some aggregate measure of the volume of the environment is necessary. While the absolute value of the volume is low at quiet times, the threshold for what can be considered "quiet" is greatly dependent on the specific microphone and phone placement. A noisy environment, however, seems to always have a high variance in volume regardless of the mean volume. Variance was therefore considered a more appropriate measure of the noisiness of the environment than the mean of the volume. The Sleep Disturbance feature is computed as the standard deviation of the audio volume data captured on all days between the hours of 12:00 AM and 06:00 AM, local time:

$$\text{Sleep Disturbance} = stdev(\text{AV}^{night}) \tag{9.2}$$

where AV^{night} represents the subset of audio volume samples collected from a participant between the hours of 12:00 AM and 6:00 AM.

Weeknight Sleep Disturbance Feature

Since people's patterns of sleep can vary between weeknights and weekends, especially in younger individuals who tend to stay up later on the weekends, we compute a second version of the Sleep Disturbance feature. The *Weeknight Sleep Disturbance* feature is defined as the standard deviation of the audio volume data captured only on *weekdays* (Sunday through Thursday) between the hours of 12:00 AM and 06:00 AM, local time:

$$\text{Weeknight Sleep Disturbance} = stdev(\text{AV}^{weeknight}) \tag{9.3}$$

where $\text{AV}^{weeknight}$ represents the subset of audio volume samples collected from a participant between the hours of 12:00 AM and 6:00 AM on weeknights only.

Summary of Measured Feature Values

Summary statistics describing the distribution of participants' feature values are provided in Table 9.1. As these features are unitless measures without any physical analog, it is challenging to provide some discussion or interpretation of their distribution. One observation which can be made is that values of the two Sleep Disturbance features are distributed very similarly, as would be expected considering that they are computed from overlapping data.

Feature	Mean (SD)	Min.	First Quartile	Second Quartile	Third Quartile	Max.
Daily Similarity	0.80 (0.07)	0.45	0.77	0.83	0.85	0.90
Sleep Disturbance	0.14 (0.06)	0.03	0.10	0.13	0.17	0.32
Weeknight Sleep Disturbance	0.14 (0.06)	0.03	0.09	0.12	0.18	0.34

Table 9.1: Summary statistics describing participants' audio volume-based features (n = 84).

9.1.3 Correlations with Mental Health Measures

To understand how well these features actually captured any symptoms of anxiety, depression, or overall poor mental health, correlations between all features and all self-report measures were computed. Table 9.2 shows the results of this correlation analysis.

Feature	Correlations with Mental Health Measures: r (P)			
	LSAS	GAD-7	PHQ-8	SDS
Daily Similarity	-0.20 (0.07)	-0.19 (0.09)	-0.37 (< 0.001)	-0.18 (0.10)
Sleep Disturbance	0.00 (0.99)	0.07 (0.52)	0.17 (0.13)	0.15 (0.17)
Weeknight Sleep Disturbance	0.05 (0.65)	0.12 (0.26)	0.23 (0.03)	0.18 (0.11)

Table 9.2: Correlations between audio volume-derived features and mental health measures (n = 84).

The Daily Similarity feature is negatively correlated with all 4 self-report measures, which supports the hypothesis that regularity in daily activity is associated with more positive mental health (i.e., lower scale scores). Daily Similarity was most strongly correlated with depressive symptoms as measured by the PHQ-8 (r = -0.37, P < 0.001). The Sleep Disturbance feature was positively correlated with scores on the GAD-7, PHQ-8, and SDS, which is in line with the hypothesis that better sleep quality (i.e., less sleep disturbance) is associated with positive mental health. The strength of the correlations are improved when only the weeknight audio is considered.

9.1.4 Discussion

To our knowledge, no other studies have inferred daily patterns of activity and inactivity solely from volume samples of ambient audio, which is captured by the Daily Similarity feature, so direct comparisons of the Daily Similarity feature with other known works are not possible. However, a feature called Circadian Movement, first proposed by Saeb et al. [44], captures the degree to which the pattern of a subject's visits to different geolocations follows a 24-hour or circadian rhythm. The study by Saeb et al. measured a significant negative correlation between Circadian Movement and PHQ-9 scores ($r = -0.63$, $P = 0.01$, $n = 18$). Our ambient volume-based measure of daily regularity in activity, called Daily Similarity, was also negatively correlated with PHQ-8 scores ($r = -0.37$, $P < 0.001$, $n = 84$). This further supports the hypothesis that regularity in daily activities is associated with lower severity of depressive symptoms. It is interesting to consider that the same underlying signal may be present in both GPS data and ambient audio data.

The association between sleep and mental health has been investigated in a number of studies. Self-reported measures of sleep quality have been shown to be associated with state anxiety [123], anxiety disorders [124], and depression [125], [126]. Sleep duration has been measured objectively in mobile sensing studies and shown to have a significant association with the severity of depressive symptoms [45], [47], [48]. Although sleep duration and sleep disturbance features are different measures and as such cannot be directly compared, these studies support our results in demonstrating a general association between sleep and depression.

9.2 Bigram Data

9.2.1 Data Overview

The 15-second audio recordings captured by the study smartphone application (captured once every 5 minutes) were processed with Automatic Speech Recognition software to detect spoken words in participants' environments. These detected words were stored as bigrams, that is, as pairs of words consisting of one word in the transcript of the audio recording and the next word immediately following it. For example, the transcript "what did you eat today" would yield the following sequence of bigrams: (what did), (did you), (you eat), (today ⟨EMPTY⟩).

Detected words were stored as bigrams, and not unigrams (single words), as we wished to have the ability to detect modification of salient words (e.g., "*not* happy" vs "*very* happy"). Since it was required that words be stored in randomized order

to prevent recreation of the transcripts, storing words as simple unigrams would not allow this. It must be made clear, however, that all analysis and feature extraction which follows only operates upon unigrams; bigrams were converted to unigrams by dropping the second word in bigram pair. Initial experimentation with bigram-based analysis was conducted but failed to achieve success beyond what was achieved in the unigram-based analysis which follows, and as a result the decision to forgo in-depth bigram-based analysis was made in order to focus efforts elsewhere.

Prior to the analysis of participants' unigram data, a threshold for sufficient data was set on the basis of the number of words detected within participants' audio recordings. Participants were included in the analysis if the total number of words detected in their ambient audio recordings was greater than a minimum of 769 words. As will be explained in the subsections which follow, a dictionary-based semantic analysis tool (LIWC) was used to analyze the semantic content of participants' unigram data. This minimum threshold was determined by noting that this semantic analysis tool was built from a corpus of words, and that the least frequently observed word category in the corpus (the sexual words category) had a mean frequency of 0.13% [127]. This implies that, on average, 1 in 769 words in the corpus fell within this category. Assuming that the word data collected from participants are similarly distributed, we would require an expected value of 769 words to detect one word in this category; hence, 769 words was the minimum threshold.

The application of this threshold for inclusion in analysis resulted in a sample of 86 participants. Within the 86-participant sample, the mean number of audio recordings captured was 3647 (SD 802), and the mean number of recordings that contained speech was 579 (SD 257). On average, 16% of recorded ambient audio contained intelligible speech. This low percentage is reasonable given that recordings were performed throughout all hours of the day. The average number of detected environmental words per participant was 4379 (SD 2625). While the original transcripts were destroyed after generation, the total number of recordings that contained detected speech was recorded for each participant. For recordings which were found to contain speech, the mean number of detected words was 7.4. This seems reasonable given that the audio recordings were 15 seconds long. All summary statistics for the total number of recordings captured, number of recordings found to contain speech, total detected words, and average word length of the transcripts, all on a per-subject basis, are presented in Table 9.3.

Statistic	Mean (SD)	Min.	First Quartile	Second Quartile	Third Quartile	Max.
Total Recordings Captured	3646 (802)	330	3764	3908	4001	4271
Recordings containing speech	579 (257)	91	391	574	740725	1288
Total detected words	4379 (2625)	841	2470	3842	5720	14882
Average number of words in recordings with speech detected	7.4 (2.01)	3.7	6.2	6.8	8.0	15.5

Table 9.3: Summary statistics for word counts of the transcripts of environmental audio recordings, per subject (n=86)

9.2.2 Analysis

A software tool, Linguistic Inquiry and Word Count (LIWC; version 2015) was used to analyze participants' words along a number of linguistic and psychological dimensions [128]. LIWC is a tool which was developed to categorize words according to both their linguistic function (i.e., which part of speech a word is functioning as a noun, adverb, etc) and according to the words' meanings with respect to psychologically-relevant concepts such as emotions, social concerns, and other constructs. Some of these categories are organized hierarchically, for example, the affect category contains the subcategories of positive and negative emotion, and the negative emotion category is further broken down into anxiety, sadness, and anger. Examples of these psychological categories, and some of the words within, are given in Table 9.4.

Category	Example words
Personal pronouns	I, them, her
Common verbs	eat, come, carry
Positive emotion	love, nice, sweet
Social Processes	mate, talk, they
Death	bury, coffin, kill

Table 9.4: Sample of Linguistic Inquiry and Word Count word categories.

Participants' environmental words were analyzed using all possible LIWC categories except summary dimensions, punctuation marks, and informal language. This resulted in 67 total categories that were tested, including the top-level categories of function words (i.e., parts of speech), other grammar (i.e., more parts of speech), affect, social, cognitive processes, perceptual processes, biological processes, drives, time orientation, relativity, and personal concerns. For each category, LIWC returns the percentage of total words analyzed which were found to belong to that specific category.

9.2.3 Feature Design

Since the LIWC categories were built with the intent to capture and count the presence of psychologically-relevant words, word counts across these different categories of words may offer insight into participants' mental health. For example, usage rates of positive emotion and negative emotion words, which are both word categories counted by LIWC, are of obvious interest. Therefore, the word detection rates in each of the 67 word analysis categories used in LIWC are each used as features to infer participants' mental health state. As such, 67 features are extracted from each participant's unigram data.

To understand how well these features capture aspects of mental health, Pearson correlations were computed between these features and each of the four self-report measures of mental health. Due to the large number of features, we wish to concisely highlight potentially interesting associations from the large number of correlations measured; therefore, only correlations with an associated P value less than .05 are presented. However, due to the large number of comparisons being performed (4 scales $\times$ 67 word categories = 268 comparisons), we considered a result statistically significant at a Bonferroni-corrected significance level of α=.0002. This is to done to counter the increase in false discoveries associated with performing multiple comparisons [129].

9.2.4 Correlations with Mental Health Measures

Table 9.5 presents the correlations between word counts of the LIWC word categories and each of the 4 self-report measures (LSAS, GAD-7, PHQ-8, and SDS) whose P values were less than .05. Please note that all tested correlations are presented in Table E.1 within Appendix E.

Of the correlations presented in Table 9.5, only the correlation between the death category and PHQ-8 scores was statistically significant at a Bonferroni-corrected significance level of α=.0002. This positive correlation shows that higher rates of death-related words detected in the environment are associated with stronger self-reported symptoms of depression.

Interestingly, the rates of words detected in the positive emotion and negative emotion categories were both measured as having very low associations with all self-report measures, with the absolute value of the Pearson r measured under 0.2 in all cases. The rates of words detected in the negative emotion category were most strongly correlated with the PHQ-8 (r=0.15, P=.17). The rates of words detected in the positive emotion category were also most strongly correlated with the PHQ-8

Mental Health Measure	Word Category	Percentage of Words: mean (SD)	Correlation with Mental Health Measure: r (P)
LSAS	death	0.16 (0.10)	0.32 (0.002)
	home	0.45 (0.14)	-0.31 (0.003)
	see	1.26 (0.28)	0.31 (0.003)
	sexual	0.22 (0.29)	-0.24 (0.024)
GAD-7	reward	1.61 (0.30)	-0.29 (0.007)
	death	0.16 (0.10)	0.27 (0.012)
	friend	0.35 (0.15)	0.26 (0.016)
	prep	11.75 (1.10)	0.24 (0.029)
	bio	2.07 (0.59)	-0.23 (0.036)
	relativ	13.57 (1.10)	-0.22 (0.045)
PHQ-8	death	0.16 (0.10)	0.41 (< 0.001)
	function	55.31 (3.13)	0.24 (0.025)
	home	0.45 (0.14)	-0.24 (0.028)
	reward	1.61 (0.30)	-0.22 (0.043)
SDS	death	0.16 (0.10)	0.28 (0.009)
	friend	0.35 (0.15)	0.24 (0.026)
	negate	2.29 (0.52)	0.23 (0.033)

Table 9.5: Top correlations between LIWC categories and LSAS, GAD-7, PHQ-8, and SDS scores (n = 86).

(r=–0.18, P=.09). Correlations and P values for all associations, including word rates in the positive emotion and negative emotion categories, are presented in Table E.1 within Appendix E.

9.2.5 Discussion

Key Findings

A key finding is the correlation between the rates of detected words within the category of death and all self-reported measures. This correlation was positive in all cases, meaning participants who had a higher proportion of death-related words detected in their ambient audio displayed worse self-reported symptoms of social anxiety, generalized anxiety, depression, and mental health-related functional impairment. The association between the use of death-related words and depression is in line with previous studies [130], [131] showing that depressed individuals tend to use more death-related words. It is important to note that these prior studies analyzed only words that were spoken or written by participants, whereas we included all the words detected in the participants' environments.

Other Interesting Findings

In light of the fact that only the correlation between rates of death-related words and the PHQ-8 was statistically significant, it is important to note that the Bonferroni correction is known to be conservative and can cause important relationships to be deemed nonsignificant [132]. That being said, this work has also revealed other interesting potential relationships between different environmental words and mental health.

The first was the positive correlation between vison-related words (the see category, including words such as "view," "saw," and "seen") and self-reported symptoms of social anxiety (r=0.31, P=.003). Higher rates of these words being associated with worse symptoms of social anxiety may be related to a known feature of the disorder. Specifically, individuals with social anxiety disorder fear the scrutiny of others, and socially anxious individuals will attempt to detect this scrutiny by visually attending to the others, especially the faces of others [133]. It may be that individuals verbalize this concern about observing this scrutiny throughout their days.

Another interesting relationship was the negative correlation between the rates of reward-related words in the environment and self-reported symptoms of generalized anxiety (r=–0.29, P=.007) and depression (r=–0.22, P=.045). Lower rates of words in this category, such as "take," "prize," and "benefit" were associated with stronger symptoms of generalized anxiety and depression. In the case of depression, this observed association may be linked to the known deficit in reward processing, and therefore, low hedonic tone noted in depressed individuals [134], [135]. If the rates of reward-related words can be used as a proxy for reward-seeking, then lower usage rates of reward-related works might be a result of this diminished capacity to focus or search out and respond to rewards. The link between reward and anxiety is less well-understood, but Gray and McNaughton [136] posit that a key feature of anxiety is related to failure or loss of reward. In this sense, anxious individuals may avoid reward-seeking to avoid triggering anxiety related to potential loss of reward. Again, if rates of reward-related words can be used as a proxy for reward seeking, this may shed some light on the observed relationship between reward-related words and symptoms of generalized anxiety.

Ambient Versus Participant-Only Word Analysis

A key feature of the methodology employed in this work is that the environmental audio recorded for each participant contained speech from any speaker in the environment — the participants themselves but also other humans and recordings (eg,

television, radio, music, etc). To the best of our knowledge, no other studies have performed linguistic analysis of audio transcripts containing speech from all ambient sources. This is important to keep in mind when we discuss previous studies that focus only upon speech or writing produced by the participant.

To provide some insight into the impact of other voices in the ambient audio and this work, it is useful to first have an estimate of how much ambient speech is typically produced by the participant and how much comes from other sources. One study [137], which employed a similar audio recording technology (with wrist-worn smart watches), determined that, of all the detected speech in the environment, roughly 18% was produced by the participant, another 18% came from other present people, and 54% from TV and radio. While the presence of other sources of speech in the audio, and therefore in the transcripts, is a confounding factor, it may also contain relevant information. While other individuals will be thought of as polluting the data, the individuals with whom one chooses to associate with may influence one's own state of mind and mental health, especially with regard to depression [138]. Similarly, the presence of words produced by TV or other media in the environmental audio could be a confound but may also contain useful information. As with the company they keep, participants' choices of media may be reflective of their state of mind and mental health. For instance, one study [139] of film preference and mental health showed an association between preference for film noire movies and depression. The presence of words spoken by individuals other than the participant may therefore still give relevant insight into the participant's state of mind and mental health.

Comparisons With Other Studies

The most reported association between participant-only word usage rates and mental health in the literature is the association between the use of first-person personal pronouns and depression. A meta-analysis [140] estimated the correlation to be small (r=0.13, 95% CI 0.10-0.16). This correlation was also measured to be quite weak in our study of ambient speech (r=0.11, P=.30) but with weaker confidence due to a much smaller sample size.

Several studies [141], [142] have investigated associations between participant-only linguistic content in social media posts and self-reported measures of anxiety and depression; these same studies have also used LIWC in their analyses and so can be compared with our work. The comparison has the caveat that our work explored speech from other parties in addition to the participant. A linguistic analysis of Facebook posts revealed positive correlations between the sadness word category and self-reported anxiety (r=0.34, P<.01) [141], whereas our study measured the corre-

lation to be much weaker (r=0.07, P=.51). They also measured the correlation between the sadness word category and self-reported symptoms of depression (r=0.22, P<.01) [141], which corresponds more closely to our results (r=0.17, P=.13). Another linguistic analysis of Facebook data also found the sadness LIWC word category to be a significant predictor of depression diagnosis (standardized regression coefficient β=0.17, P<.001) [142].

Limitations

One technical limitation of this study was the sampling technique used to capture ambient audio. Ambient audio recordings were produced quite frequently, once every 6.7 minutes (on average), but for a short duration (only 15 seconds). The short duration of recording helps to preserve smartphone battery life, but it is likely that some conversations or utterances were not captured in full. A more sophisticated sampling technique would record for a variable duration, extending the recording window until silence was detected, so that complete conversations or utterances were captured.

A fundamental limitation is due to the manner in which the environmental audio is used to generate transcripts. Automatic speech recognition software does not perform as well as human transcribers for audio recorded in noisy environments or for audio containing multiple speakers who may be interrupting one another. Furthermore, this software is often being updated and improved; therefore, reproducibility and the ability to do direct comparisons is a key concern for future studies. While this limitation is significant, it is important to also note that the accuracy of Google's Speech-to-Text API (which was used in this study) has been evaluated in clinical talk-therapy settings and demonstrating 83% sensitivity and 83% positive predictive value in detecting death-related words [143], which implies acceptable validity for the use of this type of data in our analysis.

A final limitation is related to the use of LIWC to perform the linguistic analysis of the transcripts of environmental audio. LIWC is a dictionary-based tool, and as such, categorizes words without looking at contextual information that is key to human language, ignoring sarcasm, metaphor, and analogy.

9.3 Voice Activity Data

9.3.1 Data Overview

Voice Activity is one of three audio-based streams of data extracted from audio recordings of participants' environments. This data is a binary measure of the presence or

absence of English-speaking voices in the environmental audio recordings of participants. This data was sampled with a nominal sampling period of 5 minutes, but the effective sampling period achieved by the application was, on aggregate, longer than this (the 75th percentile of average sampling periods of voice activity data was 6.4 minutes).

As with all data streams, a threshold on the minimum number of samples per participant for inclusion in analysis was set in order to handle missing data. This threshold is identical to the threshold used in the analysis of the Activity Recognition data; participants are required to have at least half the nominal number of samples expected (2,016 samples). Applying this threshold results in a sample size of 84 participants for analysis.

9.3.2 Analysis and Feature Design

The hypothesis which drove the creation of features derived from voice activity data was that individuals suffering from poor mental health would engage in fewer social interactions. Taking the presence of speaking voices in the environments of participants as a proxy measure of social interaction, features were designed to quantify social interaction. This voice activity data can only be viewed as a *proxy* for social interaction because the speaking voices detected in the environment may not correspond to the owner of the smartphone or to any person at all, in the case of voices produced by media (e.g., television, radio, etc).

The link between social interaction and social anxiety is likely the strongest among all the mental disorders considered by this work. The fear and anxiety experienced by socially anxious individuals when in social situations often causes these individuals to avoid these situations [21]. It follows, then, that a proxy measure of social interaction is likely to be predictive of symptoms of social anxiety. While avoidance of social interaction is not a hallmark feature of depression, the lack of energy and activity which does characterize depression [17] may also result in a measurable decrease in sociability.

Speech Presence Ratio **Feature**

The *Speech Presence Ratio* (SPR) feature is a proxy measure of a participant's social interaction. It is computed as the fraction of all voice activity data samples which were found to contain speaking voices:

$$\text{Speech Presence Ratio} = \frac{|\text{VA}_{voices=present}|}{|\text{VA}|} \tag{9.4}$$

where VA is the set of voice activity samples collected from a participant.

This feature was was inspired by the *Conversation Frequency* feature first described by Wang et al. in [47]. This feature differs from the Conversation Frequency feature in that the Conversation Frequency feature excluded conversations which occurred during lecture hours of the participants (which were all students) or were captured while participants were located in a lecture hall.

***Work-Time SPR* Feature**

The *Work-Time SPR* feature attempts to offer a more fine-grained focus on the times in which we would most likely expect participants to be speaking with coworkers. We are motivated to investigate this subset of the voice activity data since it is more likely to capture interactions in the workplace, which are another key cause for fear and anxiety in socially anxious individuals [17].

This feature is computed in the same manner as the Speech Presence Ratio feature, but using only the voice activity samples captured between 9 AM and 5 PM on weekdays:

$$\text{Work-Time SPR} = \frac{|\text{VA}^{worktime}_{voices=present}|}{|\text{VA}^{worktime}|} \tag{9.5}$$

where $\text{VA}^{worktime}$ is the subset of voice activity samples collected from a participant between 9 AM and 5 PM on weekdays.

***Free-Time SPR* Feature**

For the sake of completeness, the *Free-Time SPR* feature is also defined. This feature is computed in the same manner as the Work-Time SPR feature, but using only voice activity samples captured *outside* common work hours (i.e., 9 AM - 5 PM on weekdays):

$$\text{Free-Time SPR} = \frac{|\text{VA}^{freetime}_{voices=present}|}{|\text{VA}^{freetime}|} \tag{9.6}$$

where $\text{VA}^{freetime}$ is the subset of voice activity samples collected from a participant outside common work hours.

***Out-of-Home SPR* Feature**

The *Out-of-Home SPR* feature attempts to offer a more fine-grained focus on the times in which we would most likely expect participants to be engaging in social interactions

with individuals outside of the home. We are motivated to investigate this subset of the voice activity data since social interactions which occur at the home are likely occurring with people that the participants are very close with. Social interactions with close friends and family are less likely to be anxiety provoking, and therefore not likely trigger any urges to avoid these interactions.

This feature is computed in the same manner as the Speech Presence Ratio feature, but using only voice activity samples captured while the participant was not at their home:

$$\text{Out-of-Home SPR} = \frac{|\text{VA}^{!home}_{voices=present}|}{|\text{VA}^{!home}|} \tag{9.7}$$

where $\text{VA}^{!home}$ is the subset of voice activity samples collected from a participant when they were outside of their home. A Participant's clustered GPS location data is used to determine the times at which they are at home, using the method described in Section 7.2.2.

At-Home SPR Feature

For the sake of completeness, the *At-Home SPR* feature is also defined. This feature is computed in the same manner as the Out-of-Home SPR feature, but using only voice activity samples captured when participants are at their home:

$$\text{At-Home SPR} = \frac{|\text{VA}^{home}_{voices=present}|}{|\text{VA}^{home}|} \tag{9.8}$$

where VA^{home} is the subset of voice activity samples collected from a participant when they were at the home.

Summary of Measured Feature Values

Summary statistics describing the distribution of participants' feature values are provided in Table 9.6. Please note that the features were computed for two different subsets of participants, as follows. The Speech Presence Ratio and the Work-Time and Free-Time SPR features were all computed for the subset of participants who had at least 50% of expected audio samples (n = 84). The Out-of-Home and At-Home SPR features were computed for only 71 of these 84 participants, since these features additionally rely upon location data and only 71 participants had sufficient audio and location data.

Inspection of the basic Speech Presence Ratio feature reveals that participants were in the presence of English-speaking voices 15% of the time, on average. This appears

Feature	Mean (SD)	Min.	First Quartile	Second Quartile	Third Quartile	Max.
VAD Ratio	0.15 (0.06)	0.01	0.11	0.15	0.20	0.30
Work-Time VAD Ratio	0.21 (0.11)	0.01	0.12	0.20	0.28	0.48
Free-Time VAD Ratio	0.13 (0.06)	0.01	0.09	0.13	0.17	0.30
Out-of-Home VAD Ratio	0.25 (0.12)	0.00	0.15	0.23	0.32	0.69
At-Home VAD Ratio	0.12 (0.06)	0.01	0.08	0.12	0.16	0.28

Table 9.6: Summary statistics describing participants' voice activity-derived features (n = 84 for the VAD Ratio, Work-Time VAD Ratio, and Free-Time VAD Ratio; n = 71 for the Out-of-Home and At-Home VAD Ratios)

to align with intuition, as these voice activity samples were taken at all hours of the study, including nighttime. Focusing on the Work-Time and Free-Time SPR features, participants encountered more speech during work-time hours (21% of the time, on average) than they did outside these hours (13% of the time, on average). This also is reasonable, considering that a sizeable portion of the data collected during "Free-Time" would contain late nights where we would expect participants to be sleeping. Similarly, participants encountered more speech when out of the home (25% of the time, on average), versus when at home (12% of the time, on average).

9.3.3 Correlations with Mental Health Measures

To understand how well these features actually captured any symptoms of anxiety, depression, or overall poor mental health, correlations between all features and all self-report measures were computed. Table 9.7 shows the results of this correlation analysis.

Feature	Correlations with Mental Health Measures: r (P)			
	LSAS	GAD-7	PHQ-8	SDS
VAD Ratio	-0.19 (0.081)	-0.16 (0.143)	-0.37 (0.001)	-0.29 (0.008)
Work-Time VAD Ratio	-0.11 (0.320)	-0.15 (0.172)	-0.26 (0.015)	-0.19 (0.086)
Free-Time VAD Ratio	-0.19 (0.084)	-0.12 (0.266)	-0.34 (0.001)	-0.27 (0.012)
Out-of-Home VAD Ratio	0.20 (0.089)	0.01 (0.939)	0.09 (0.475)	0.05 (0.709)
At-Home VAD Ratio	-0.29 (0.015)	-0.13 (0.280)	-0.35 (0.003)	-0.25 (0.036)

Table 9.7: Correlations between voice activity-derived features and mental health measures (n = 84 for the VAD Ratio, Work-Time and Free-Time VAD Ratios; n = 71 for the Out-of-Home and At-Home VAD Ratios).

The Speech Presence Ratio feature was found to be negatively correlated with all self-reported measures of mental health, and significantly correlated (at a 5% significance level) with depression as measured by the PHQ-8 (r = -0.37, P = 0.001) and general impairment as measured by the SDS (r = -0.29, P = 0.008). The correlations between mental health measures and the Work-Time, Free-Time, and At-Home SPR features all followed this same trend, but were generally weaker in almost all cases.

The Out-of-Home SPR feature breaks this trend, showing positive correlations across the board with all mental health measures.

9.3.4 Discussion

The negative correlation between the Speech Presence Ratio and all mental health measures indicates that participants who spent less time in the presence of speech had worse mental health. This finding aligns with our hypothesis. The significant correlation between the Speech Presence Ratio and PHQ-8 scores (r = -0.37, P = 0.001) also replicates a previous finding. The pioneering study by Wang et al. [47] inferred the number of conversations that subjects encountered throughout their days by performing an analysis of ambient audio and also found a significant negative correlation with PHQ-9 scores (r = -0.39, P = 0.02, n = 48). The specific feature designed by Wang et al. is more contextually aware than our Speech Presence Ratio feature because not all audio that they found to contain speech was considered conversational in their system. Specifically, they ignored speech if it was detected during lecture or group meeting hours of their student subject population. Despite the fact that our proxy measure of social interaction is context unaware, we measured similar results. Time spent proximal to human speech was also found to be significantly associated with changes in PHQ-9 scores (P=.048) in a study by Ben-Zeev et al. [48].

A curious finding was that the direction of correlation between the Out-of-Home SPR feature and mental health measures is *positive*. While the correlations between Out-of-Home SPR and the GAD-7, PHQ-8, and SDS are all non-significant and very weak, the correlation with the LSAS is nearing significance (r = 0.20, P = 0.089). This positive correlation with the LSAS, which is a measure of social anxiety, may indicate that socially anxious individuals who engage in more social interaction outside of the home have stronger symptoms of social anxiety. The original hypothesis was that more socially anxious individuals would *avoid* social interaction, and therefore demonstrate less social interaction. Such a relationship would therefore produce a negative correlation between social engagement and LSAS scores. The measured *positive* correlation might indicate that some socially anxious individuals may be *unable* to avoid social situations, which would therefore trigger their anxiety and elevate their LSAS scores. In this scenario, when avoidance is not an option, more social interaction is associated with worse symptoms of social anxiety, which is in line with the noted characteristics of social anxiety [17].

Limitations

The Speech Presence Ratio feature and its variants do not distinguish between recorded speech (e.g., from a TV or radio) and human speech — they simply indicate the presence of intelligible speech. This method does not distinguish between speakers, so in many cases, the participant themselves may not be the person speaking. The method also only detects English speech, so it will potentially miss speech if, for example, a participant does not speak English at home. Finally, our technique for detecting speech using automatic speech recognition is likely to be more biased toward false negatives than false positives. That is, that if speech is detected by the system, it is highly likely that the speech is present, yet it is much more likely to miss speech in noisy environments or environments with multiple speakers speaking concurrently.

9.4 Conclusion

This chapter has described the design of features derived from audio recordings of participants' environments. Analysis of the *volume* of these audio recordings led to the design of features which quantify regularity in participants' daily activity and disturbances to participants' sleep, both of which were shown to be correlated with symptoms of depression and general impairment. All features derived from audio volume were novel designs. Analysis of the *presence of speaking voices* in these audio recordings replicated previous findings which measured a correlation between time spent proximal to speech and severity of depression. The baseline Speech Presence Ratio was adapted from prior work, while the four variants thereof are novel contributions. Finally, analysis of the *words* detected in the audio recordings revealed several relationships, including the use of death-related words, reward-related word, and words related to vision being potentially associated with self-reported measures of social anxiety, generalized anxiety, depression, and general psychiatric impairment. To the best of our knowledge, this work is the first to perform linguistic analysis of environmental or ambient audio recordings captured by individuals' smartphones.

This chapter is the last of three chapters to focus upon the design of features which were intended to be predictive of anxiety and depression. The next chapter, Chapter 10, will select a subset of these features and test their ability to predict clinical levels of anxiety and depression.

Chapter 10

Screening for Anxiety and Depression

Chapters 7 through 9 have described the engineering of features which were were designed to give insight into individuals' state of mind with respect to symptoms of anxiety and depression. The degree to which these features did indeed associate with symptoms of anxiety and depression was tested by measuring correlations between feature values, which were extracted solely from smartphone-collected data, and study participants' self-reported mental health scale data. This section provides a more in-depth test of the ability of features to predict mental health, by using a subset of the features to build predictive models of study participants' mental health. Specifically, these models will be used *screen* participants for social anxiety disorder, generalized anxiety disorder, and depression, by predicting if participants scored above or below diagnostic thresholds on the respective self-report measures.

This chapter is organized as follows. The next section, Section 10.1 ("Data Preparation"), describes the process used to select a subset of features for use in the predictive modelling. This section also describes the process by which participants' *known* screening results were determined prior to prediction. Section 10.2 ("Screening Model Development and Performance") describes an experiment where different classes of predictive models were built for each prediction task, and compares and contrasts their relative predictive performance. In Section 10.3, one high-performing class of model is selected for in-depth inspection and interpretation. Finally, the chapter is concluded in Section 10.4.

10.1 Data Preparation

10.1.1 Feature Selection

A total of 97 features were implemented and extracted from participants' smartphone-collected data, 67 of which were extracted from the bigram data stream alone. Considering that this number of features is roughly on par with the total number of participants recruited in the study (112), this is a relatively large number of features. Many of these features were shown to have poor associations with mental health measures, and of those that did exhibit moderate or strong associations, many features were measured to be correlated with one another and therefore would be redundant to include. The aim of the feature selection process was to select a small number of well-performing features for use in predictive modelling. The desired number of features was in the range of 8-10, in order to maintain a rough 10:1 ratio of samples (i.e., participants) to independent variables (i.e., features). While not a strict requirement, this is in-line with commonly cited rules of thumb for avoiding over-fitting [144].

Candidate features were selected from each data stream by selecting those which had the strongest broad correlations across the multiple mental health measures yet were weakly correlated with other features from the same data stream. This yielded an initial set of 11 candidate features. Figure 10.1 contains all pairwise correlations between the 11 candidate features. Table 10.1 presents the correlations between the 11 candidate features and all self-report mental health measures.

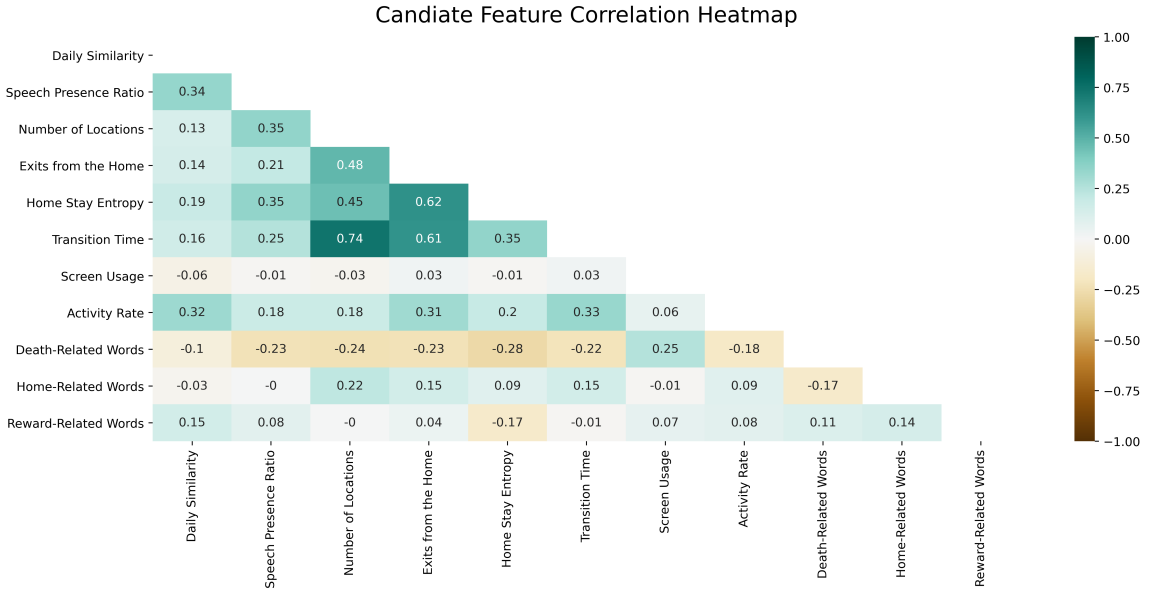


Figure 10.1: Correlations between candidate features used in model building and evaluation.

Feature	Correlations with Mental Health Measures: r (P)			
	LSAS	GAD-7	PHQ-8	SDS
Daily Similarity	-0.16 (0.183)	-0.18 (0.149)	-0.37 (0.002)	-0.18 (0.135)
Speech Presence Ratio	-0.19 (0.115)	-0.15 (0.216)	-0.41 (0.001)	-0.29 (0.018)
Number of Locations	-0.24 (0.048)	-0.24 (0.049)	-0.30 (0.012)	-0.27 (0.027)
Exits from the Home	-0.21 (0.085)	-0.27 (0.023)	-0.26 (0.035)	-0.21 (0.092)
Home Stay Entropy	-0.26 (0.034)	-0.22 (0.074)	-0.31 (0.009)	-0.20 (0.097)
Transition Time	-0.25 (0.040)	-0.27 (0.029)	-0.32 (0.008)	-0.32 (0.008)
Screen Usage	0.14 (0.245)	0.07 (0.559)	0.13 (0.274)	0.31 (0.011)
Activity Rate	-0.26 (0.034)	-0.06 (0.613)	-0.28 (0.019)	-0.18 (0.141)
Death-Related Words	0.34 (0.005)	0.20 (0.098)	0.36 (0.003)	0.24 (0.050)
Home-Related Words	-0.33 (0.007)	-0.19 (0.117)	-0.36 (0.002)	-0.12 (0.318)
Reward-Related Words	-0.18 (0.150)	-0.31 (0.011)	-0.24 (0.054)	-0.10 (0.410)

Table 10.1: Correlations between candidate features and mental health measures (n = 80 participants).

The candidate set was further pruned down by eliminating those features which were correlated with other candidates with $|r| > 0.4$, which resulted in a final feature set of 8 features. Participants which had insufficient data such that they were missing half or more of these features were dropped from the data set; this yielded a subset of 80 participants for analysis. Table 10.2 presents the final 8 features selected for analysis, including summary statistics describing the distribution of the feature values.

Feature	Mean (SD)	Min.	First Quartile	Second Quartile	Third Quartile	Max.
Daily Similarity	0.80 (0.07)	0.45	0.77	0.83	0.85	0.90
Speech Presence Ratio	0.15 (0.06)	0.01	0.11	0.15	0.20	0.30
Number of Locations	15.2 (9.93)	1.00	8.00	13.00	20.50	50.00
Screen Usage	0.23 (0.13)	0.02	0.12	0.21	0.29	0.60
Activity Rate	8.70 (5.26)	0.57	5.01	7.98	10.54	25.04
Death-Related Words	0.17 (0.10)	0.00	0.09	0.15	0.20	0.51
Home-Related Words	0.45 (0.14)	0.15	0.34	0.44	0.54	0.83
Reward-Related Words	1.61 (0.27)	0.90	1.43	1.61	1.81	2.28

Table 10.2: Summary statistics describing the distribution of features in the final feature set (n = 80)

It must be noted that one criticism of this feature selection technique is that it is not bias-free. The selection of features on the basis of their correlations with mental health measures (measured using the full set of participants) introduces bias since features which are already known to correlate well with participants' mental health measures will naturally perform well in predicting which participants are above or below threshold on the mental health measures used. This criticism is valid, yet there are some arguments to be made which show that this bias is small enough to be acceptable for this work. Namely, that a number of the selected features were

already shown to be known good indicators of mental health in prior works: the Speech Presence Ratio, Number of Locations, and Screen Usage features. Since these features were already shown to have significant associations with mental health in other samples of individuals, the likelihood that they are overfit to our particular sample is lower than it would be otherwise. Furthermore, it must be noted that this work is *exploratory* in nature, and that some degree of bias is acceptable since any conclusions drawn must be subject to further confirmatory analysis.

10.1.2 Participant Mental Health Screening

The modelling reported in this chapter used participants' features to predict whether they would screen positively for social anxiety disorder, generalized anxiety disorder, or depression. In order to measure the accuracy of the predictions, it is necessary to know if participants actually did screen positively for these disorders. Participants were screened for social anxiety disorder, generalized anxiety disorder, and depression, using the LSAS, GAD-7, and PHQ-8 instruments, respectively. The screening criteria and the number and percentage of positive screenings are summarized in Table 10.3. The positive screening rates for all three disorders within this subset of 80 participants were notably high with respect to rates in the general population; a discussion on this general finding was provided in Section 6.3 where the screening rates for the entire sample of 112 participants were discussed.

Disorder	Screening Criterion	Positive Screenings
Social Anxiety Disorder	LSAS score $\geq$ 60	29 (36%)
Generalized Anxiety Disorder	GAD-7 score $\geq$ 10	20 (25%)
Major Depressive Disorder	PHQ-8 score $\geq$ 10	28 (35%)

Table 10.3: Participant screening results for social anxiety disorder, generalized anxiety disorder, and depression (n = 80)

10.2 Screening Model Development and Performance

A variety of different machine learning models were used to predict participants' screening results. This section describes the model building and evaluation process, and presents and compares the predictive performance of the different models.

10.2.1 Models Classes

A total of 11 different classes of models were built and evaluated. All of the models were built and evaluated using the software library mlr3 [145], which is a machine

learning framework for the R programming language. All of these models, once trained, return a single screening prediction given a participant's features as input. Since the models return a single output, independent models were built to screen for social anxiety disorder, generalized anxiety disorder, and depression. The following classes of models were used:

1. Classification Tree

   Classification Tree [146] models were built using the Classification Tree Learner from the mlr3 learners package [147]. Classification trees predict a discrete class label (i.e., a positive or negative screening) given a set of features by performing a series of comparisons between each feature value and a threshold value (learned during training) that best separates the classes.

2. Linear Discriminant Analysis (LDA)

   Linear Discriminant Analysis [148] models were built using the Linear Discriminant Analysis Classification Learner from the mlr3 learners package [149]. LDA classifiers predict a class label by learning the linear combination of feature values which best separates the two classes. This linear combination of features reduces the n-dimensional feature set to a single new dimension, and compares the resulting value in this new dimension to a internal threshold acting as the decision criterion (this internal threshold, like the linear combination, is also learned during training).

3. Logistic Regression

   Logistic Regression [150] models were built using the Logistic Regression Classification Learner from the mlr3 learners package [151]. Logistic regression computes a weighted sum of feature values, which are then passed through a sigmoidal activation function, to ultimately output the probability of the given set of features belonging to the positive class (e.g., the probability that, given a set of feature values, that the individual associated with those features would screen positive for a given disorder). The values of the weights (sometimes referred to as coefficients) used for the computation of the weighted sum are learned during training. When used in classification tasks, the probability output by the logistic regression model is compared to a threshold probability (e.g., 50%) to assign a class label to the given feature values.

4. Logistic Regression with LASSO regularization

   Logistic Regression with LASSO regularization [152] models were built using the Generalized Linear Model with Elastic Net Regularization Classification Learner

from the mlr3 learners package [153]. LASSO regularization changes the manner in which the optimal weights, or coefficients, are learned in the training phase of a logistic regression model. This modification often results in some feature coefficients dropping to zero, which has the effect of de-selecting certain features. This can be useful, in practice, for reducing overfitting, which produces models with better predictive performance on unseen or out-of-sample data.

5. Naive Bayes

   Naive Bayes [154] models were built using the Naive Bayes Classification Learner from the mlr3 learners package [155]. A Naive Bayes classifier computes the posterior probability that a given set of feature values belong to a particular class, with the key simplifying assumption that features are independent (i.e., that the probability of a positive screening given one feature is not influenced by the values of any of the other features). The training phase estimates the likelihood functions for each feature in the negative and positive classes. Given a set of learned likelihood functions and a set of feature values, the posterior probabilities of the features belonging to the negative or positive classes are computed. A class label is then made on the basis of which posterior probability is higher.

6. Random Forest - Standard Trees

   Random Forest [156] models were built using the Ranger Classification Learner from the mlr3 learners package [157]. A random forest model is a group (or *ensemble*) of many differently-built classification trees. Each classification tree makes its own prediction, and the prediction made by the random forest is formed by noting which prediction result was made most often the trees (in a sense, the constituent trees "vote").

   Two distinct classes of random forests were used, which differed in the maximum allowed depth for the trees in the forest (the depth of a tree is the number of sequential comparisons made between feature values and thresholds). In this class, no limit was placed on the maximum depth of the trees.

7. Random Forest - Stumps Only

   This Random Forest class limits the maximum depth of the trees to 1 (trees of depth 1 are often referred to as "stumps").

8. Support Vector Machine with Linear Kernel

   Support Vector Machine (SVM) [154] models were built using the Support Vector Machine Learner from the mlr3 learners package [158]. SVMs classify data by

learning the optimal decision line (or hyper-plane, in higher-dimensional spaces) that best separates the classes on the basis of feature values. SVMs are therefore naturally linear classifiers, but they can effectively perform non-linear classification by first making a non-linear mapping of features to a new space prior to classification, using what is referred to as a *kernel*.

Four distinct classes of SVMs were used, which differed in the kernel used. In this class, a linear kernel was used, to perform standard linear classification.

9. Support Vector Machine with Polynomial Kernel

   This SVM class was set to use a polynomial kernel.

10. Support Vector Machine with Radial Basis Function Kernel

    This SVM class was set to use a radial basis function (RBF) kernel.

11. Support Vector Machine with Sigmoid Kernel

    This SVM class was set to use a sigmoid kernel.

The complete set of model settings and hyperparameter values for all models are provided in the code listing in Appendix F.

10.2.2 Model Training and Performance Evaluation

Models were trained and tested using a nested cross-validation resampling strategy [159] to estimate model performance in an unbiased manner (i.e., performance is measured using data that was not used to train the model) . This strategy consists of a 2-level cross-validation, with one cross-validation loop nested within another (see Figure 10.2).

In the outer loop of the resampling, the dataset (features and true screenings) is first split into training and test sets, called the *outer* training sets and the *outer* test sets. In the inner loop, each outer training set is further broken down into *inner* training and *inner* test sets. Models are repeatedly trained on the inner training sets with varying hyperparameter values, and the hyperparameter values which yield the highest average performance on the inner test sets are identified and recorded. Then, moving back up to the outer level, each model is then re-trained on the full outer training set, and its performance is evaluated on the outer test set using the optimal hyperparameter values discovered in the inner loop iterations.

Figure 10.2, which is reproduced from [160], provides an example illustration of such a nested cross-validation strategy, using 3-fold cross-validation in the outer loop and 4-fold cross-validation in the inner loop. In this example, 3 final models would

have been evaluated, and performance of the overall model would be estimated as the *average* of the 3 model's individual performance scores.

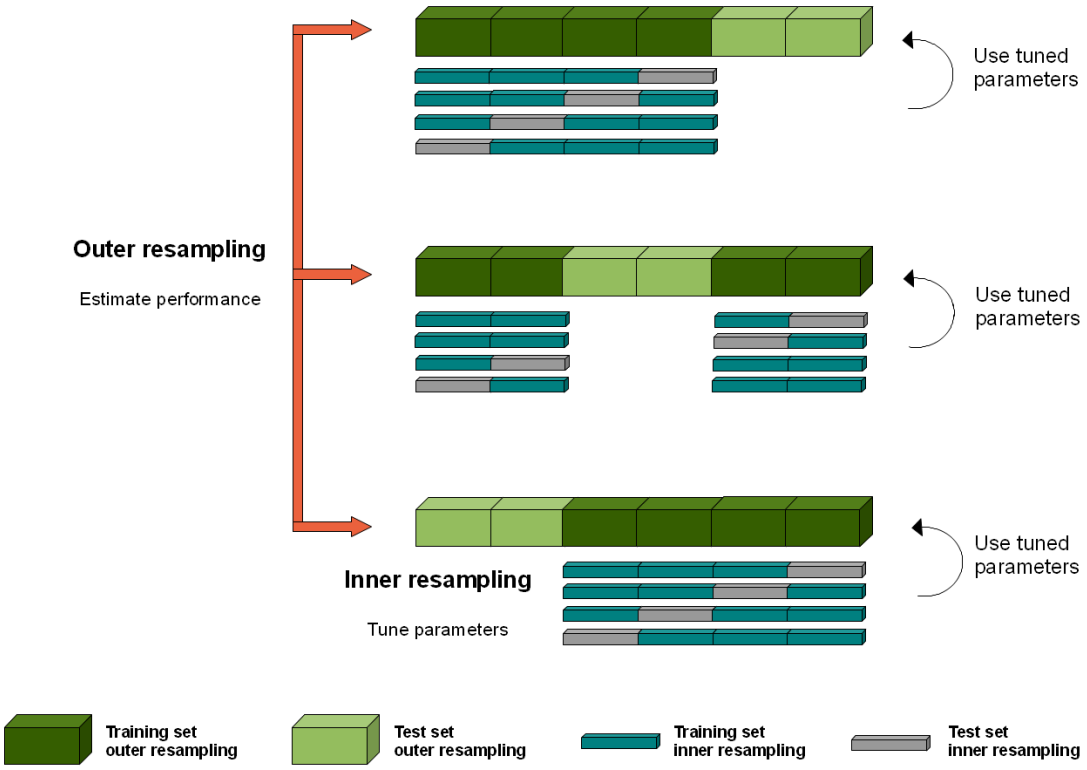


Figure 10.2: Visualization of a nested cross-validation resampling strategy [160]

The resampling strategy used in this work was performed with 5-fold cross validation in both the outer and inner loops. Furthermore, this process was repeated 20 times, each time with a different random assignment of data to the folds. This resampling strategy therefore produced 100 models for each class of model and each target disorder. In total, 3,300 models were evaluated (11 classes of models; 3 target disorders; 100 resampling repetitions). The performance metric used to compare model performance was the Area Under the Receiver Operating Characteristic (AUROC, sometimes referred to as AUC - "Area Under the Curve") [161]. The code listing in Appendix F also contains this nested cross-validation resampling procedure.

This repeated k-fold cross validation scheme was selected instead of a more basic train and test split (where, for example, 80% of data is selected for training and the remaining 20% of data is used for testing) because of the small dataset size and relatively high variability in the data. Model performance can vary quite drastically depending one which set of participants are used for training and which set are used for testing. Instead of biasing the results by picked one high-performing split, this

approach performs many random splits and aggregates the results to provide a less biased result. In the results which follow, both the mean and standard deviation of the models' performance will be presented, which will help to illustrate this variance in performance.

To test whether the mean predictive performance of each class of models is significantly better than an uninformative model (which would produce random predictions), a modified version of a paired samples T-test was used. This corrected repeated k-fold cross validation t-test was developed by Bouckaert and Frank in [162] to account for the fact that the performance of each model produced by a repeated k-fold cross validation resampling is not independent, since models built using this strategy share training data.

10.2.3 Results

The mean performance (and additionally the standard deviation) of each model class in predicting social anxiety disorder, generalized anxiety disorder, and depression is given in Tables 10.4, 10.5, and 10.6 respectively, and summarized in Figure 10.3. Also plotted in each chart in Figure 10.3 is a dotted horizontal line at a mean AUROC of 0.5. This level corresponds to the performance which would be achieved by an uninformative model which makes random predictions, and is being used as a baseline comparison for model performance.

Focusing first upon the prediction of social anxiety disorder, the measured performance values for each model class are presented in Table 10.4. Model performance values range from a mean AUROC of 0.43 (SVM with RBF kernel) to 0.67 (the LDA, Logistic Regression, and Naive Bayes models). Note that an AUROC of less than 0.5 is possible when assessing model performance on out-of-sample (i.e., test) data, as is the case here, and this indicates that the associated models were likely overfit to their training data. Seven of the 11 model classes achieved a mean performance which was significantly greater than an uninformative model (at a 5% significance level): the LDA, Logistic Regression, Naive Bayes, both Random Forest model classes, and the SVM models with Linear and Sigmoid kernels.

Shifting focus to the prediction of generalized anxiety disorder, the measured performance values are presented in Table 10.5. Model performance values range from a mean AUROC of 0.44 (SVM with Sigmoid kernel) to 0.56 (LDA). None of the 11 model classes achieved a mean performance which was significantly greater than an uninformative model (at a 5% significance level).

Finally, the measured performance values in predicting depression are presented in Table 10.6. Model performance values range from a mean AUROC of 0.53 (Logistic

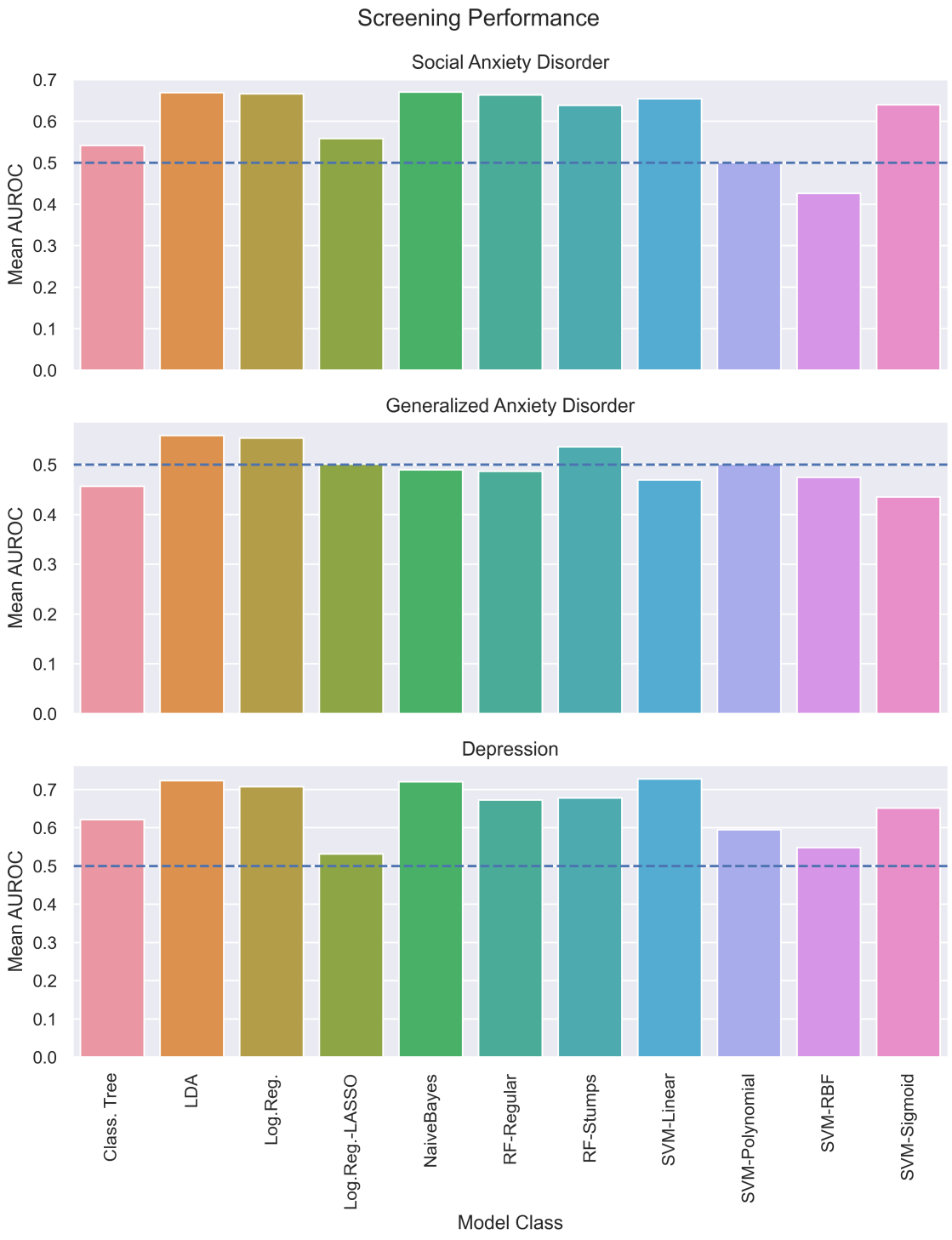


Figure 10.3: Screening performance for each model class and disorder.

Regression with LASSO regularization) to 0.73 (SVM with a Linear kernel). Eight of the 11 model classes achieved a mean performance which was significantly greater than an uninformative model (at a 5% significance level).

Model Class	Mean AUROC (SD)	t	p
Class. Tree	0.54 (0.14)	0.600	0.275
LDA	0.67 (0.11)	2.932	0.002
Log.Reg.	0.67 (0.12)	2.746	0.004
Log.Reg.-LASSO	0.56 (0.10)	1.130	0.131
NaiveBayes	0.67 (0.11)	3.088	0.001
RF-Regular	0.66 (0.12)	2.686	0.004
RF-Stumps	0.64 (0.12)	2.185	0.016
SVM-Linear	0.65 (0.16)	1.930	0.028
SVM-Polynomial	0.50 (0.15)	0.007	0.497
SVM-RBF	0.43 (0.13)	-1.109	0.135
SVM-Sigmoid	0.64 (0.15)	1.842	0.034

Table 10.4: Social anxiety disorder screening results for all model classes.

Model Class	AUROC: Mean (SD)	t	p
Class. Tree	0.46 (0.11)	-0.747	0.228
LDA	0.56 (0.15)	0.780	0.219
Log.Reg.	0.55 (0.16)	0.671	0.252
Log.Reg.-LASSO	0.50 (0.01)	0.196	0.422
NaiveBayes	0.49 (0.15)	-0.135	0.446
RF-Regular	0.49 (0.13)	-0.191	0.424
RF-Stumps	0.54 (0.14)	0.508	0.306
SVM-Linear	0.47 (0.14)	-0.437	0.332
SVM-Polynomial	0.50 (0.18)	0.012	0.495
SVM-RBF	0.47 (0.13)	-0.373	0.355
SVM-Sigmoid	0.44 (0.15)	-0.848	0.199

Table 10.5: Generalized anxiety disorder screening results for all model classes.

10.2.4 Discussion

A key finding when comparing prediction results between the three disorders is that, on aggregate, the prediction of depression was the most accurate. The prediction of social anxiety disorder achieved worse performance, on aggregate, than depression, yet some models were still found to have significant predictive capability. Predictive models of generalized anxiety disorder achieved the lowest accuracy, however, and none of the eleven classes of models investigated exhibited performance significantly greater than a baseline uninformative (random prediction) model.

The success of the predictive models of social anxiety disorder and depression is interesting when contrasted with the failures of the generalized anxiety disorder models. The particular feature set used in this work seems unable to predict generalized anxiety disorder. This may be related to how generalized anxiety manifests itself in individuals' feelings and behaviors in a different manner than social anxiety or depression. Individuals suffering from social anxiety disorder commonly employ behavioral avoidance to avoid situations which trigger their anxiety symptoms – they may choose specific jobs to avoid public speaking, or avoid going to social gatherings [163]. The

Model Class	Mean AUROC (SD)	t	p
Class. Tree	0.62 (0.11)	2.204	0.015
LDA	0.72 (0.11)	3.817	< 0.001
Log.Reg.	0.71 (0.12)	3.343	0.001
Log.Reg.-LASSO	0.53 (0.08)	0.780	0.219
NaiveBayes	0.72 (0.12)	3.621	< 0.001
RF-Regular	0.67 (0.11)	3.021	0.002
RF-Stumps	0.68 (0.12)	2.959	0.002
SVM-Linear	0.73 (0.12)	3.845	< 0.001
SVM-Polynomial	0.59 (0.17)	1.094	0.138
SVM-RBF	0.55 (0.19)	0.494	0.311
SVM-Sigmoid	0.65 (0.14)	2.058	0.021

Table 10.6: Depression screening results for all model classes.

relative lack of or decrease in public speaking, travelling, and leaving the home are all physical behaviors which can be detected by some of the features used in this study: the Speech Presence Ratio feature can measure changes in the amount of speech, the Number of Locations feature can measure changes in amount of travel, and the Activity Rate feature directly measures physical activity. While depressed individuals do not avoid specific situations for the same reasons that socially anxious individuals do, depression is characterized by a general lack of motivation and energy [164]. This lack of motivation and energy also appears to manifest itself in behaviors (or lack thereof) which are detectible by our feature set (specifically the Speech Presence Ratio and Locations Visited features).

This behavioral avoidance can be contrasted with the cognitive avoidance that is common in individuals who suffer from GAD [165], which includes maladaptive and somewhat pathological strategies like distraction, worry, and thought suppression [166]. These are all strategies used by individuals who suffer from GAD to suppress their symptoms, yet they do not as easily manifest in the physical realm in a way that can be detected through, for example, GPS location data. In other words, it may be that this study's feature set does a good job of detecting behaviors, but a poor job in inferring the cognitive strategies which may often be used in a pathological manner to combat negative or aversive feelings. While it is true that screen usage-based features could infer times where subjects were employing distraction as a behavioral avoidance strategy, there are other possible explanations for screen time that might not include avoidance. The screen usage data which we collected did not contain information on which apps were being used while the screen was on, which could be used to know the motivation for the use of the phone. This more fine-grained usage data which detailed which apps were in use (e.g. games versus productivity apps such as email, etc) would yield more insight into this type of avoidance and

therefore be may more predictive of GAD.

10.2.5 Comparisons With Other Studies

A number of studies have used smartphone-collected data in a similar fashion to build and evaluate predictive models of depression. Saeb et al. achieved a classification accuracy of 78.8% in detecting depressed individuals (which was defined as having a PHQ-9 score ≥ 5) using only location-based features [44]. To enable more direct comparisons, two existing studies were found which also reported AUROC as their metric of accuracy in predicting depression using smartphone-collected data. Wang et al. [53] report an AUROC of 0.81 for a model that predicted depression which used location, audio, and screen-based features. A study by Place et al. [167] achieved an AUROC of 0.74 for detecting depressed mood using features derived from GPS location, audio, motion sensor data, phone and messaging metadata, screen data, and other device data.

Both the studies by Wang et al. [53] and Place et al. [167] differ from ours, however, in how they screen subjects for depression, as they used scores from the abbreviated 2-item PHQ-2 depression instrument. Furthermore, in both studies participants were drawn from a more homogenous sample, as they were both geolocated in the particular metro areas. Participants in the study by Wang et al. [53] were 48 Dartmouth College undergraduate students, while the 73 participants in the study by Place et al. [167] were all residents of the Boston area. Participants in our study were a mix of students and non-students, and were geographically located across Canada in both rural and urban areas. The nature of the area in which subjects live and travel in is of particular import for location-based features, as individuals living in rural areas may exhibit different patterns of travel than those in urban areas.

Studies of anxiety disorders using smartphone-collected data are much less represented in the literature. A study by Boukechba et al. [168] demonstrated strong correlations between smartphone-collected data and symptom severity of social anxiety disorder, but no classification (i.e. predictive screening) was performed. To our knowledge, there are currently no studies that have predicted or otherwise measured correlations between generalized anxiety disorder and smartphone-collected data.

10.3 Logistic Regression Model Interpretation

The LDA, Logistic Regression, Naive Bayes, and Random Forest model classes were consistently the top-performing classes of predictive models across all disorders (see Figure 10.3). Given that logistic regression models are frequently encountered in the

medical literature concerning diagnostic modelling [169], and that they have a easily-interpretable structure, this section offers an analysis of these models to understand how the individual features contributed to the screening predictions made by the models.

In order to inspect and interpret the models and the determine relative influence of each feature within a model, a *single* model is trained using the entire data set for each disorder under investigation. This is in contrast with the approach used in Section 10.2, where a resampling strategy was used to build multiple models in order to estimate model performance on unseen data. The coefficients of these full models will then be presented and discussed. These coefficients are x-standardized as all features are all scaled to have zero mean and standard deviation of one prior to model building. This standardization is necessary since we wish to compare the effects of the features which do not share a common unit of measurement or scale.

10.3.1 Results

The standardized logistic regression coefficients of the models of social anxiety disorder (SAD) and depression (MDD) are presented in Figure 10.4. The model of generalized anxiety disorder is not included since it failed to achieve significant predictive capability.

First, by looking at the signs of the feature coefficients, we can determine which features were found to decrease the odds of a positive screening, since negative coefficients correspond to a feature which is associated with decreased odds of a positive screening (a characteristic of the logistic regression modelling used). The following features were associated with decreased odds of a positive screening for both social anxiety and depression: Activity Rate, Daily Similarity, Home-Related Words, Number of Locations, and Reward-Related Words. This indicates that these features are generally capturing healthy behaviours, with respect to social anxiety and depression. By contrast, the Death-Related Words and Screen Usage features were both associated with increased odds of positive screening of social anxiety disorder and depression.

The Speech Presence Ratio feature was found to have contrasting directionality with respect to screening for social anxiety disorder and depression. A greater Speech Presence Ratio was associated with greatly decreased odds of a depression screening but was also associated with a slight increase of odds of a social anxiety disorder screening.

Secondly, in terms of the relative magnitude of their associated coefficients, several features stand out. The Death-Related Words feature was the largest risk factor for

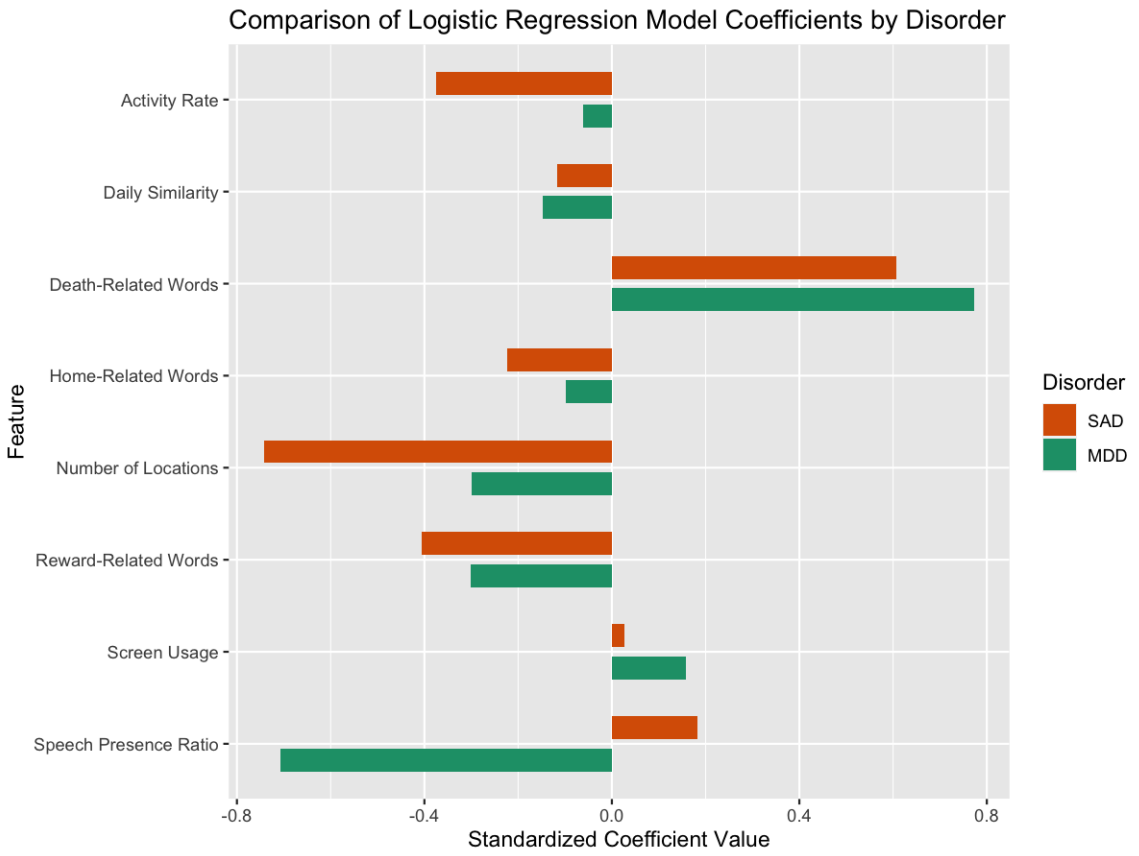


Figure 10.4: Comparison of logistic regression model coefficients by disorder.

both social anxiety disorder and depression. The greatest protective factor (i.e., the feature associated with the largest *reduction* in odds for a positive screening) for social anxiety disorder was the Number of Locations feature. The greatest preventative factor for depression, however, was the Speech Presence Ratio factor.

10.3.2 Discussion

Comparing the models of SAD and MDD, the proportion of Death-Related Words detected in the environmental audio recordings is the greatest risk factor for both disorders. The link between the usage of death-related words and depression has been demonstrated in some previous works [130], [131], but we are aware of no empirical studies that have demonstrated a link between death-related words and symptoms of social anxiety disorder. Some researchers have proposed that death anxiety and the fear of death may function as a transdiagnostic construct underpinning a range of mental disorders, including anxiety disorders [170]. If this hypothesis holds true, the presence of death-related words in environmental audio may also be serving as a proxy measure of death anxiety.

Continuing to compare the models of SAD and MDD, the Activity Rate, Daily Similarity, and Number of Locations features all appeared as protective factors for both disorders. These results are all in line with the respective hypothesis that drove the creation of these features. Namely, that more physical activity (Section 7.1.2), greater regularity in the pattern of activity (Section 9.1.2), and more mobility/less avoidance (Section 7.2.2) would all be associated with better mental health. Greater proportions of Home-Related Words and Reward-Related words detected in participants' ambient audio recordings were also protective factors for both social anxiety disorder and depression. As was discussed in Section 9.2.5, the use of more reward-related words might indicate greater reward seeking behavior and higher hedonic tone, which could be associated with decreased anxiety and weaker symptoms of depression, respectively. To speculate on the Home-Related Words result, it may be the case that individuals with sub-threshold symptoms of social anxiety and depression use home-related words more often because they find their home to be a healthy, positive place.

Contrasting the two models, the Speech Presence Ratio feature operates in opposite directions in the models of SAD and MDD. It was a protective factor with respect to depression: subjects who spend more time in the presence of speech in their environment had greatly reduced odds of depression. The same was not true of social anxiety disorder, in which the direction of this effect was reversed. This result that links more environmental speech to weaker symptoms of depression is one that replicates results from prior studies [47], [48]. That this relationship does not also hold for social anxiety disorder is interesting. It may be the case that, once adjusting for factors like social avoidance (Number of Locations feature), more speech is associated with stronger symptoms of social anxiety. This seems plausible given that interactions with strangers can be a trigger for symptoms of social anxiety.

10.4 Conclusion

This chapter has described how a subset of the features designed in Chapters 7 through 9 were selected and used to build models which screened participants for social anxiety disorder, generalized anxiety disorder, and depression. A variety of different models were built and benchmarked, and multiple models were found which could screen participants for social anxiety disorder and depression with performance that was significantly better than uninformative models. The best-performing models screened for social anxiety disorder and depression with a mean AUROC of 0.67 and 0.73, respectively. One class of predictive model used — logistic regression — was then

further analyzed in order to determine how features influenced the models' outputs and to compare and contrast the disorders in terms of their model structure. This chapter culminates the work done to infer mental health from smartphone-collected data. The next chapter provides a summary of the thesis and proposes future work for research in this space.

Chapter 11

Conclusion

In this work, we have shown that objective data, passively collected from individuals' smartphones, gives broad insight into individuals' behaviors and activities. Furthermore, it was shown that some of these patterns of behaviors and activities are associated with mental health to such a degree that it is possible to use them to screen individuals for social anxiety disorder and depression.

While the ability to screen individuals for anxiety and depression is an achievement in and of itself, this work has also taken key steps towards the overarching goal of objectively quantifying and tracking mental health. While this goal is lofty and seemingly distant, the challenges which were identified, studied, and solved in this narrower formulation of the problem — objective screening for anxiety and depression — are likely to also appear when pursuing the more general goal of the objective quantification of broad mental health. The identification of these key challenges and the methods developed to solve them, therefore, are all contributions to further research in the field which are highlighted in this chapter.

11.1 Contributions

11.1.1 Exploration of Willingness to Consent to Smartphone-based Assessment of Mental Health

Given the potential for mobile technology to monitor patients' mental health symptoms in a passive and pervasive way, it is helpful for researchers to understand how patients may respond to requests for access to their personal data. While some may see it as a foregone conclusion that individuals would consent to this, by noting that mental health-related apps already exist and have user bases, it is however not clear how many individuals in the general population would choose *not* to use such

a system. One contribution of this work, described in Chapter 3, was a study which surveyed a patient population for their willingness to consent to allowing smartphone applications access their personal data for the purpose of tracking their mental health. Key findings of this study were that 84% of respondents indicated partial or complete willingness to install a mental health-monitoring application. Willingness to provide data collection across different sources ranges from 18% to 46%, with more intrusive or private sources of data being more likely to be withheld. These findings are not restricted to mental health apps targeting any specific disorders, and as such offer insight to the broad community of mental health researchers. We were among the first to ask this kind of question, and others have followed suit, as it is an important precursor to projects in the direction of this thesis. As of this printing, the research article which first published these results [171] has been cited 29 times.

11.1.2 Passive Smartphone Data Collection System

Having identified the potential for smartphones to act as mobile sensing platforms, and having also confirmed individuals' willingness to engage with such technology, a complete smartphone-based data collection system was built. This system, described in Chapter 5, handles all steps in the data collection procedure from participant enrollment, collection of both objective digital data and self-report mental health measures, and secure storage of data. The objective digital data collected and stored by this system could be employed in similar studies of other mental disorders by simply exchanging the embedded self-report measures of anxiety and depression to alternative measures as needed.

11.1.3 Remote, Anonymous, and Automated Study Design

Chapter 4 of this work contributes a general methodology which allows for data collection from human participants in a manner that is remote, anonymous, and automated. The recruitment procedure, which utilizes an online recruitment platform, can easily be scaled to recruit participants in much larger sizes than what was shown in this work. This scalability is achieved through the combination of two necessary components. Firstly, the existence of a large population of willing study participants, which is offered by the Prolific online recruitment platform. Equally as important, however, is the ability of the data collection system to dynamically adapt to increasing numbers of new participants *without* requiring research staff to monitor the system and manually provision new computing resources as necessary. This study design, in conjunction with the complete data collection system, enables researchers to easily

recruit large samples of participants without face-to-face contact, large numbers of support staff, or computer system administrators.

11.1.4 Hypothesis-Driven Feature Engineering

One of the key goals of this work was the extraction of patterns from participants' objective smartphone-collected data which were predictive of their symptoms of anxiety and depression. To this end, the diagnostic criteria and characteristics of these disorders were studied and used to drive the design of novel features. While prior works have not been oblivious to disorders' characteristics when designing their features, the links between disorders and hypothesized features have often been implicit or sometimes lacking in references to theoretical foundations in the psychological and/or psychiatric literature.

One contribution of this work has been the clear progression from disorder symptomatology to hypotheses on how these symptoms might manifest in smartphone-collected data. These hypotheses were then used to build mathematical formulations of features which were designed to capture manifestations of these symptoms. Finally, these features were extracted and associations between these features and self-reported measures of anxiety and depression were discussed. This approach has yielded a number of novel features which were found to have associations with social anxiety and depression, included features derived from the volume of sounds in participants' environments. Not only are features such as Daily Similarity novel, they have been extracted from a data stream — ambient audio volume — which have never been analyzed in any prior work in this space. The Daily Similarity feature is one of the many features designed in this fashion and presented in Chapters 7 through 9.

11.1.5 Predictive Modelling of Anxiety and Depression

Finally, to assess our progress towards the goal of building automated and objective measures of anxiety and depression, several key features were used to build models which predict participants' screenings of social anxiety disorder, generalized anxiety disorder, and depression. This predictive modelling, presented in Chapter 10, makes a number of contributions. Firstly, it adds to prior work which has attempted predict depression from smartphone-collected data, studying the same problem on a new sample of participants and showing similar results in the prediction of depression. Secondly, it is, to our knowledge, the first study which has attempted to predict generalized anxiety disorder using smartphone-collected data. Finally, it is the first study which has compared and contrasted how the same set of 8 features can be used

to predict these three separate disorders. The analysis of this comparative modelling has provided a number of insights which we believe will be important for future work in this space, especially works which will attempt to model, predict, or measure generalized anxiety disorder using smartphone-collected data.

11.2 Limitations

This section will present, at a high level, some of the threats to the validity of the research and other limitations that should be addressed in future works.

The first general limitation of this work is centered upon the manner in which features are used to infer relevant behaviors. While the link between behaviors and mental state are generally well known and supported by the literature (by design), the link between smartphone-collected data and behavior is less strong. In this work, behaviors are inferred from digital data through feature engineering, but the inferred behaviors have not been *confirmed* as being accurately representative of actual behaviors exhibited by participants. Consider, for example, the inference of the number of locations visited from GPS data. What if, for example, an individual chose to leave the home without bringing their smartphone with them? In such an instance, the inferred behavior of the individual — no travel and no locations visited — would not be truly representative of their actual behavior. Additionally, the analysis of ambient spoken word is also impacted by the presence of words not uttered by the participant under study (see Sections 9.2.5 and Sections 9.3.4 for in-depth discussions of the limitations with bigram and voice activity data, respectively). One method by which the validity of inferred behaviors could be assessed would be to conduct studies where participants are asked to actively log their actual behaviors of interest, and then compare inferred behaviors to actual behaviors to assess their validity.

A second limitation of this work is the reliance upon self-reported mental health data. Self-reported data is practical from an operational perspective, but offers a less accurate assessment of mental state than clinician-administered measures. The problem of inaccurate metrics of mental health state are not entirely solved by transitioning to clinician-administered tools. Indeed, this work is in part motivated by the imperfection of standard clinical assessment. As such, this problem is somewhat of a Catch-22, where we are motivated to produce better metrics, but in order to assess the accuracy of any new metric, it is necessary to compare to imperfect gold standards. Any metric which is a more accurate measure of mental health than the gold standard will, for at least some individuals, produce scores which deviate from the standard (by necessity), and this deviation will present as inaccuracy when com-

paring these new scores to those produced by the gold standard. One solution to this paradox would be to forgo trying to replicate the output produced by a gold standard in an observational study (such as this one), but instead determine if any new candidate metric can detect response to treatment sooner than the gold standard in a study which employs some intervention already proven to improve mental health.

A final limitation of this work is related to how its findings may be used in the future to design interventions. One may conclude, for example, that since participants in this study which traveled to more locations had weaker self-reported symptoms of depression, that one intervention to treat depression is simply to make depressed individuals travel more. Due to the purely observational nature of this study, however, one cannot conclude that the correlation between the number of locations visited and depression severity also implies a causal link between mobility and depression. A more appropriate study design to test the efficacy of such an intervention would be a randomized control trial, where extrapolation of causality from results can be made [172].

11.3 Future Work

There remain a number of different areas in which research towards objective measures of mental health can be advanced. This section provides a discussion on some of the these areas of exploration.

11.3.1 New Data Streams and Features

Research in this space, including this work, has not in any way exhaustively explored the broad set of sources of digital data which may offer insight into individuals' mental health. While further confirmatory studies will be necessary to establish a set of reliably predictive features of anxiety and depression, exploratory studies which seek to explore new data streams and new features should also be pursued.

One data stream which offers a wealth of potential insight into mental state is language, whether spoken or written (i.e., text messages or chats). This work has performed some lexical analysis on spoken language, but with the emergence of modern natural language processing techniques, much more complex analyses of human language are possible [173]. This could also be combined with speaker identification techniques [174] which can distinguish between different speaking voices. This combination could allow future analyses to not only gain better insight into spoken language in participants' environments, but to also gain crucial contextual information about which individual is speaking at any time, modelling relationship dynamics in individ-

uals' social lives. In addition to the words spoken and typed to other individuals, the words typed into computer systems such as search engines can also be analyzed in a similar fashion, as this data is likely to also contain relevant information.

Another key set of data which was not used in this work is physiological data. Physiological data such as heart rate, body temperature, blood pressure, and electrocardiogram (ECG) signals can be collected by wearable smart devices with accuracy close to that of professional high-end sensors [175]. This data is immediately promising for the study of anxiety disorders, as prior work has shown some efficacy in assessing anxiety symptoms from physiological data [176]. The combined use of features which capture physiological state with features which capture higher-level behaviors and context (such as the features used in this work) may offer greater accuracy in inferring mental health state. As an example of some of the potential synergies in combining these classes of features, consider the study of social anxiety disorder. Knowing that an individual is out of the home and interacting with a stranger, as indicated by behavioral features like Speech Presence Ratio and Exits from the Home, this may be an ideal time to inspect physiological data for the presence of physiological stress or arousal.

11.3.2 Different Disorders and Transdiagnostic Features

Mental health is a broad space, of which anxiety disorders and depression are only a small part. While research using passive sensing methodologies is underway in the study of different disorders, including substance abuse [177], post-traumatic stress disorder [167], bipolar disorder [178], and schizophrenia [179], these research efforts are typically siloed to their respective disorders. It is, of course, necessary for these studies to focus on a single disorder, given the complexity of the task and the large number of unknowns.

Moving forward, however, studies which can apply the passive sensing methodology to the simultaneous study of symptoms of *multiple* disorders will be very valuable. Such studies may be able to uncover the presence of *transdiagnostic* features, features which capture behaviors, states, or processes which underlie *multiple* disorders [180]. These transdiagnostic features could offer new ways of classifying, measuring, and treating mental disorders that are not rooted in the traditional taxonomy of mental disorders. Given the known and prominent comorbidity between many classes of disorders, including mood and anxiety disorders [181], it is likely that there are observable links between these disorders that could be captured by transdiagnostic features. A better understanding of these transdiagnostic features and how they relate to the disorders may offer potential new avenues for understanding and treating

comorbid disorders.

This work has taken a first step towards this, by simultaneously modelling social anxiety disorder, generalized anxiety disorder, and depression (see Section 10.3). We hope that this approach will be adopted by other researchers in exploring broad associations between many other classes of disorders.

11.3.3 Supporting New Mobile Software and Hardware

Android-based smartphones were the sole mobile sensing platform used in this work. Going forward, the ability to use smartphones as passive and persistent sensing platforms is at risk, given the trend of vendors in limiting background processing and preventing applications from accessing personal data. Alternative mobile software and hardware may be necessary to collect this same data in the same manner as this work, in the face of these restrictions. One alternative is the use of wearable devices such as fitness bands and smart watches. These devices typically have battery-efficient hardware for passive sensor data collection, specifically with regards to audio and motion sensor data. The key challenge will be to identify which devices allow developers to write custom software for those devices, such that they can access raw data in a programmable and flexible way. Some smart wearable device vendors essentially act as data custodians, only allowing developers to access user data through their own web-based API, which does not grant free flexibility in data access.

11.3.4 Personalized Modelling

The modelling performed in this work operates on a group level, where information extracted from all study participants is used to classify or label individuals' mental state. The general associations that exist within the broad group therefore are used to make inferences about individuals.

A challenge in this approach is that the associations between features and outcomes, say between Locations Visited and depression, may differ among individuals. While on aggregate there may be a negative association between locations visited and depression severity in the entire group, it may be the case that for some small subgroup of individuals, their particular set of mental and physical characteristics cause this particular relationship to differ in strength or even in direction. This subgroup of individuals may be small enough not to exert a significant influence on the association when measured with the entire group, yet the broad group-level association is not representative of their particular subgroup's, and therefore the group-level model does not characterize them well.

One approach that could improve predictive performance for such outliers would be to identify these subgroups of outliers, and build models specifically for these subgroups [182]. A new problem that emerges, then, is the identification of these particular subgroups. An alternate approach is to transition towards individual models personalized to a single person (i.e., subgroups of size 1). A broad, group-level model could act as a "draft" model for an individual. Then, during an initial recalibration phase, this model could be personalized to an individual by observing a sufficient quantity of their data over time to learn their baseline personal dynamics, in terms of how their features relate to their mental health. This re-training or re-fitting procedure may rely on weeks or months of observational data from an individual. Going forward, the personalized model could be used to detect changes from baseline mental states. This approach could then, for example, detect when a healthy individual may be developing a disorder or relapsing, if their personalized model was built during a healthy period of their life. Alternatively, a personalized model may be able to detect remission in individuals who have been living with a disorder for some time, if their personalized model was built during a period in which they were suffering from a disorder.

11.3.5 Enhanced Privacy

Future work in this space, in any direction, will certainly require larger sample sizes of participants. One impediment to the recruitment of participants is likely the concern that the deeply personal and private data collected throughout these studies may become leaked to hackers or bad actors. Alternatively, even without leaks to individuals outside the research staff, participants might object to researchers themselves having the ability to inspect their personal data. There are some avenues for the enhancement of participant privacy that were not utilized in this work, but will likely become necessary in the future.

Firstly, we note that, in this work and many prior works, participants' *raw* data are extracted from their smartphones and stored in centralized databases for later feature extraction. This approach is excellent for experimentation, since having access to the raw data (e.g., all location samples collected across time) enables researchers to continually experiment with the development of new features derived from that data. An alternative approach would be to determine ahead of time which features are to be extracted from participants' data, and have the software running on participants' smartphones extract these features from their raw data. In this manner, only the feature values are collected and stored by researchers, and not the raw sensor data, which contains more information and therefore represents more of a privacy risk. As

the research in this field matures and researchers land upon some commonly-used features, such an approach may be go part way towards alleviating privacy concerns.

The recording of features only, and not the raw sensor data, still does not eliminate risks to participant privacy. These features are still, by design, indicative of individuals mental state, and therefore represent patient information. An additional safeguard to maintain privacy and build trust could be the adoption of an approach called *differential privacy* [183]. This emerging approach allows for the maintenance of a database of patient information that allows for aggregate or group-level analyses, without offering any information on individuals. For example, a database which maintains differential privacy allows researchers to study associations between features and self-reported mental health states, or use these features to build group-level screening or diagnostic models, all without allowing researchers to specifically inspect any one individual's feature values or mental health scores. While individual analysis is obviously necessary for something like a patient medical record or for individualized modelling, group-level analysis — which is what is performed by many studies — may suffice.

11.4 Conclusion

This chapter has concluded the thesis by summarizing the key contributions of the research and presenting some of the wide areas of opportunity for the further advancement of research in this space. We hope that this nascent area of interdisciplinary research continues to grow and gather input from many other researchers and subject matter experts from the broader worlds of engineering, technology, ethics, psychology, and psychiatry to help produce solutions which will ultimately aid those in need.

Appendix A

Self-Report Measures of Mental Health

GAD-7

Over the last 2 weeks, how often have you been bothered by the following problems? *(Use "✔" to indicate your answer)*	Not at all	Several days	More than half the days	Nearly every day
1. Feeling nervous, anxious or on edge	0	1	2	3
2. Not being able to stop or control worrying	0	1	2	3
3. Worrying too much about different things	0	1	2	3
4. Trouble relaxing	0	1	2	3
5. Being so restless that it is hard to sit still	0	1	2	3
6. Becoming easily annoyed or irritable	0	1	2	3
7. Feeling afraid as if something awful might happen	0	1	2	3

(*For office coding: Total Score T*____ = ____ + ____ + ____)

Developed by Drs. Robert L. Spitzer, Janet B.W. Williams, Kurt Kroenke and colleagues, with an educational grant from Pfizer Inc. No permission required to reproduce, translate, display or distribute.

PATIENT HEALTH QUESTIONNAIRE-8 (PHQ-8)

Over the last 2 weeks, how often have you been bothered by any of the following problems? *(Use "✔" to indicate your answer)*	**Not at all**	**Several days**	**More than half the days**	**Nearly every day**
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself — or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3

Developed by Drs. Robert L. Spitzer, Janet B.W. Williams, Kurt Kroenke and colleagues, with an educational grant from Pfizer Inc. No permission required to reproduce, translate, display or distribute.

Appendix B

Willingness Survey

RESEARCH STUDY INTEREST QUESTIONNAIRE

Here at the S.T.A.R.T. Clinic we are conducting research, along with a team of researchers from the University of Toronto, about a new way to measure social anxiety. This new method of measuring social anxiety would use a smartphone app that would be given alongside, or would entirely replace, the questionnaire-based methods that you may have already seen or filled out. This app would gather and record information from your phone to look at the activities in your day-to-day life that are relevant to a person's level of social anxiety. The app would then share some of the relevant information with your doctor to allow them to better diagnose problems and prescribe treatment. While we believe this application can help both patients and doctors, we also understand that this technology has the potential to impact peoples' privacy.

To help better understand how this technology can be developed to improve clinical care while also respecting peoples' right to privacy, we would greatly appreciate if you could answer some questions regarding this app. We are not asking you to use such an app or to enroll in any studies or trials, we are only interested in your thoughts and opinions on how you would feel about using it. Please be assured your responses will be treated as confidential, will not become part of your personnel file, and will in no way affect any treatment that you receive here at the S.T.A.R.T. clinic.

The first 3 questions are more general questions regarding your use of smartphones.

For each question, please circle your answer and provide additional information, if necessary.

1. Do you own a smartphone and use it daily?
    a. Yes
    b. Yes, but I don't use it daily.
    c. No

2. Do you connect to the Internet on your smartphone, either using a mobile data plan or WiFi?
    a. Yes
    b. No

3. Would you be willing to install and use a smartphone app to help your doctor better diagnose mental health problems and/or provide treatment?
    a. Yes.
    b. Maybe, but I would need to know more information first.
    c. No
    d. Other- specify: _____

The next series of questions will ask you how willing you would be to share certain pieces of private information with doctors and the research team through the smartphone app. Please note, that in each case, the information would only be shared with doctors here at the START Clinic and a research team from the University of Toronto for the purpose of developing the technology. The data would be stored securely and you would be able to have it deleted at any time should you so choose.

For each question, please circle your answer and provide additional information, if necessary.

4. Would you be willing to have the app collect and share your location? This would use your smartphone's GPS and would pinpoint your location on a map from time to time throughout the day.
    a. Yes, I would be willing to have the app collect and share this information.
    b. I might be willing, but would need more information from my doctor first.
    c. No, I would not be willing to have the app collect and share this information.
    d. Other- specify: _____

5. Would you be willing to have the app record the number of contacts you call or send SMS (text messages) to, and the dates & times when you phone or text them? The identities of your contacts would be not be shared.
    a. Yes, I would be willing to have the app collect and share this information.
    b. I might be willing, but would need more information from my doctor first.
    c. No, I would not be willing to have the app collect and share this information.
    d. Other- specify: _____

6. Would you be willing to have the app read the contents of your text messages to look for keywords related to mental health? For example, looking for the usage of words like "tired", "depressed", or "happy". The whole text messages would not be shared, only detected keywords.
    a. Yes, I would be willing to have the app collect and share this information.
    b. I might be willing, but would need more information from my doctor first.
    c. No, I would not be willing to have the app collect and share this information.
    d. Other- specify: _____

7. Would you be willing to have the app record every time you create a calendar entry in your calendar app? The specifics of the calendar entry or event would not be shared, only the date & time that you create or modify it.
    a. Yes, I would be willing to have the app collect and share this information.
    b. I might be willing, but would need more information from my doctor first.
    c. No, I would not be willing to have the app collect and share this information.
    d. Other- specify: _____

8. Would you be willing to have the app record every time you turn on or off your phone's screen (using the "power" or "lock" button)?
    a. Yes, I would be willing to have the app collect and share this information.
    b. I might be willing, but would need more information from my doctor first.
    c. No, I would not be willing to have the app collect and share this information.
    d. Other- specify: _____

9. Would you be willing to have the app use its sensors to try and detect if and when you are walking, running, in a car, or standing still?
    a. Yes, I would be willing to have the app collect and share this information.
    b. I might be willing, but would need more information from my doctor first.
    c. No, I would not be willing to have the app collect and share this information.
    d. Other- specify: _____

10. Would you be willing to have the app occasionally turn on and use the phone's microphone to record the sounds of your surroundings? This audio would be used by the app to try and classify your surrounding as loud, quiet, or busy, but it would not attempt to recognize any words that you or people around you speak, nor would it be listened to by humans. It would not record your phone calls, only ambient audio from time to time when you are not making a phone call
    a. Yes, I would be willing to have the app collect and share this information.
    b. I might be willing, but would need more information from my doctor first.
    c. No, I would not be willing to have the app collect and share this information.
    d. Other- specify: _____

11. Consider the same scenario as question 10, but in this case the app will also detect and recognize specific words being spoken aloud by you or anyone else present in the recording. The app would attempt to recognize specific keywords like "tired", "depressed", or "happy". This speech recognition would be done in software by the app and the audio will never be listened to by humans.
    a. Yes, I would be willing to have the app collect and share this information.
    b. I might be willing, but would need more information from my doctor first.
    c. No, I would not be willing to have the app collect and share this information.
    d. Other- specify: _____

12. Consider the same scenario as question 11, but in this case the app is also able to perform any software-based processing to the recorded audio in order detect things that may be relevant to your mental health. This processing (in whatever form) would be done in software by the app and the audio recordings will never be listened to by humans, nor will humans ever read a transcript of the recordings.
    a. Yes, I would be willing to have the app collect and share this information.
    b. I might be willing, but would need more information from my doctor first.
    c. No, I would not be willing to have the app collect and share this information.
    d. Other- specify: _____

Research Study Interest Questionnaire

The last question is a general question related to your use of smartphones. Please circle your answer and provide additional information, if necessary.

13. What type of phone do you use daily? If you use multiple phones (work and personal), please select whichever corresponds to your personal phone.
    a. iPhone
    b. Android
    c. Blackberry
    d. Windows mobile phone
    e. Other- specify: _____

Appendix C

Study Description and Consent Form

Smartphone monitoring for mental health and quality of life measurements

Hosted by *ece.utoronto.ca*

£18.00 • 120 minutes • £0.00/hr • 0/20

In this study you will be asked to install an app onto your smartphone and to let it run for 2 weeks, but you will only need to do 2 hours of active work. The app will collect information and data about where you go and what you do, without any input from you. We will require you to actively install the app, run through a short app setup, and answer a series of questionnaires in the app. The questionnaires will appear twice, once at the beginning of the study and again at the end of the study, 2 weeks after. The rest of this description contains a Consent Guide with more detail on the study and the app. Please read the Consent Guide and only choose to participate in this study if you agree to install the app, provide the app with all the necessary permissions, and answer the questionnaires.

Consent Guide
INFORMED CONSENT

You are being asked to consider participating in a research study conducted by a group of researchers from the Department of Electrical and Computer Engineering at the University of Toronto.

This guide explains the purpose of the research study, provides information about measuring anxiety and/or depression symptoms, the procedures involved, possible risks and benefits, and the rights of participants.

Please read this guide carefully. Your participation in this study is entirely voluntary - you do not have to take part. Only agree to the study if you agree to the terms of the study described in this guide. If you do not agree to any of the terms, please do not enroll in the study.

INTRODUCTION

INTRODUCTION

This study is being done because researchers want to develop an improved way to measure how a person's symptoms of anxiety and/or depression may be affecting their life. There is evidence to suggest that smartphones and smartphone applications (or "apps") can help with this because they can record a person's state of mental well-being and their activities throughout the day, when they are not in contact with their doctor.

WHY IS THIS STUDY BEING DONE?

The purpose of this study is to determine if a mental health monitoring app, like the one described above, can detect symptoms of anxiety and/or depression in the general population: people who may or may not have any mental health diagnoses and who may or may not be receiving treatment, therapy, or counselling.

WHAT WILL HAPPEN DURING THIS STUDY?

Participants will have the mobile application ("app") installed on their phone. During the study, the app will be collecting and recording information from your smartphone about your day-to-day activities and communications. The mobile application will collect the following information from your smartphone:

1. Location: The application will use your smartphone's GPS that allows it to pinpoint your location on a map from time to time throughout the day.

2. Battery: The application will record your smartphone's battery charge level from time to time throughout the day.

3. Contacts: The application will record the number of contacts you call or send SMS (text messages) to, and the dates & times when you phone or text them. The names or numbers of your contacts will not be recorded.

4. Text Messages (SMS): The application will read the contents of your text messages and software tools will be used to look for keywords related to mental health. For example, we may detect the usage of words like "tired", "depressed", or "happy". This detection of keywords is done by automated software; human beings will not read your messages, and we will not save your messages after processing them.

5. Calendar: The application will record every time you create a calendar entry in your calendar app. The specifics of the calendar entry or event would not be recorded, only the date & time that you create or modify it.

6. Screen State: The application will record every time you turn on or off your phone's screen (using the "power" or "lock" button).

7. Light Intensity: The application will use your smartphone's light sensor to measure and record the intensity (brightness) of the ambient light in your surroundings from time to time throughout the day.

8. Physical Activity: The application will use your smartphone's sensors (including the accelerometer) to detect if and when you are walking, running, in a car, or standing still.

9. Audio: The application will use your smartphone's microphone to record short durations of audio from time to time throughout the day. Only ambient audio will be recorded, and this ambient audio may pick up the sounds of you and/or people around you speaking. Phone calls will not be recorded, only ambient audio when your phone is not making a call. This audio will be processed by automated software in order detect aspects of the sound that may be relevant to your mental health. The relevant aspects that might be detected are the following:

a. The volume of the sounds in your surroundings

b. The presence of a conversation in your surroundings

c. The utterance of certain keywords, "tired", "depressed", or "happy".

This processing is done by automated software only; no humans will listen to the recordings and the audio recordings will not be saved after processing them.

The application will also ask you to answer a series of questionnaires twice. You will answer the questionnaires once when the app is installed and then again at the end of the study before you delete the app.

WHAT ARE THE RESPONSIBILITIES OF STUDY PARTICIPANTS?

You must have an Android smartphone running Android version 4.4 or greater. You must install the study app from the Google Play Store (the study URL will take you to the app page on the Play Store). You must run the app after installing it in order to set it up.

During the app setup, you must be grant the Android application the necessary permissions to operate. When you first install and run the app, the setup will ask you for a number of permissions in order to collect the data described above. If you do not grant all the permissions requested you will not be eligible for the study and you will not receive any payment.

You must be willing to respond to the study questionnaires at the beginning and end of the study.

STUDY TIMELINE AND ACTIVITIES
If you decide to participate in the study, you will be asked to do the following:

App Install

The study URL will take you to the app in the Google Play Store where you can install it. You will then install the app on your phone and you will be walked through the setup procedure where you will provide the app the necessary permissions and then enter your Prolific ID into the app. After the app is installed

and set up, you will be asked to complete the questionnaires, in the app, for the first time. Once this step is complete, the application is running and will begin collecting information in the background for the next 8 weeks.

This process will take about an hour.

App Uninstall

After the 8-week period is over, you will receive a notification on your phone asking you to complete the final part of the study. You will be asked to complete the same set of questionnaires that you did in the beginning. After completing them, you will be given the study completion code/completion URL. Once you get the completion code/URL, you can uninstall the app and the study is complete.

The process will take about 45 minutes.

WHAT ARE THE RISKS OR HARMS OF PARTICIPATING IN THIS STUDY?
You will be asked to answer the questionnaires on the phone twice. This may be challenging at times or frustrating. Furthermore, the questions may be time consuming to answer. Also, the questions themselves may be pointed and directed to your mood. As a result, they may make you annoyed, sad, or angry. Study staff will review your survey responses, and if any responses show mental distress or a risk of self harm or harm to others, we will send a message to you directing you to get help from a mental health distress hotline in your province or territory. Subjects who show mental distress or a risk of self harm or harm to others will also be removed from the study and will receive full payment for their participation.

You will also be asked to allow the smartphone application to collect sensitive and personal data about your life. We have designed automated software to process this data so that no human beings will be manually reading, looking at, or otherwise scrutinizing this personal information to limit the invasion of your privacy. Despite this, you may feel uneasy or uncomfortable about the fact that the smartphone application is collecting information about your daily activities.

WHAT ARE THE BENEFITS OF PARTICIPATING IN THIS STUDY?
There is no direct benefit to you for participating in this study. Your participation may or may not help doctors measure other people's level of anxiety and/or depression in a way that is faster, easier, or more accurate than by questionnaire in the future.

WHAT ARE THE COSTS OF PARTICPATING IN THIS STUDY?
There will be no costs to you as a result of your participation in this study.

HOW WILL MY PRIVACY BE RESPECTED?
Research staff have taken measures to limit their access to your data, meaning that they will only ever look at your data to check that the app is running and collecting data. The research staff have created software to extract the meaningful information from your data. This software allows them to conduct their research without ever having to directly look at, read, or listen to any of your data.

HOW WILL MY INFORMATION BE KEPT CONFIDENTIAL?
The investigators and researchers will keep the information they see or receive about you confidential, to the extent permitted by applicable laws. As some electronic data is stored on computers on American soil, your data could be accessed by the US government as per the PATRIOT act. Even though the risk of identifying you from the electronic data is very small, it can never be completely eliminated.

The data collected during this trial will be stored indefinitely, with no set timeline for deletion. If you withdraw from the study, you may request that your data be deleted. Your data will be retained if you do not make this request.

The data collected during this trial will be used to conduct research, which may be publicly published. None of the data we collect from you or your identity will ever be included in any research publication. You may request a summary of any research publications that arise as a result of this trial. Please contact Professor

Jonathan Rose at 416 978-6992 or at jonathan.rose@ece.utoronto.ca if you wish to be given access to such publications if they are made in the future.

PROTECTING THE DATA ON YOUR SMARTPHONE

The app used in this research study stores some of the data it collects temporarily on your phone. In order to prevent any potential breach of privacy in the event that someone gets access to your phone, we recommend that you use a security feature like a PIN code, passphrase, fingerprint, or the equivalent to prevent unauthorized access to your phone

DO THE INVESTIGATORS HAVE ANY CONFLICTS OF INTEREST?

There are no conflicts of interest to declare related to this study.

WHAT ARE MY RIGHTS AS A STUDY PARTICIPANT?

You do not waive your legal rights by participating in this study. You have the right to ask questions and receive answers throughout this study.

This study has also been reviewed by the University of Toronto Research Oversight and Compliance Office. You may also contact the University of Toronto Research Oversight and Compliance Office - Human Research Ethics Program if you have questions about your rights as a participant:

- Be E-mail: ethics.review@utoronto.ca
- By telephone: 416-946-3273

QUALITY ASSURANCE

The research study you are participating in may be reviewed for quality assurance to make sure that the required laws and guidelines are followed. If chosen, (a) representative(s) of the University of Toronto Human Research Ethics Program (HREP) may access study-related data and/or consent materials as part of the

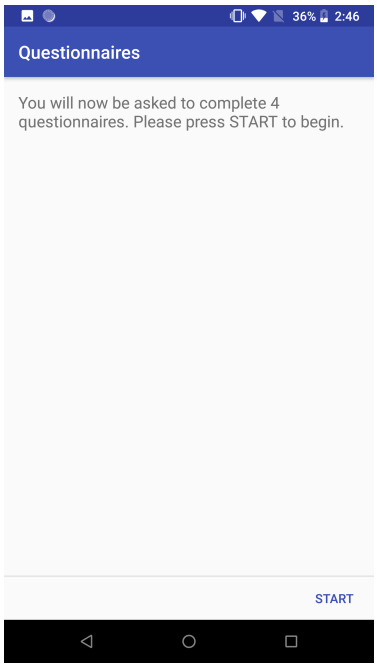
review. All information accessed by the HREP will be upheld to the same level of confidentiality that has been stated by the research team.

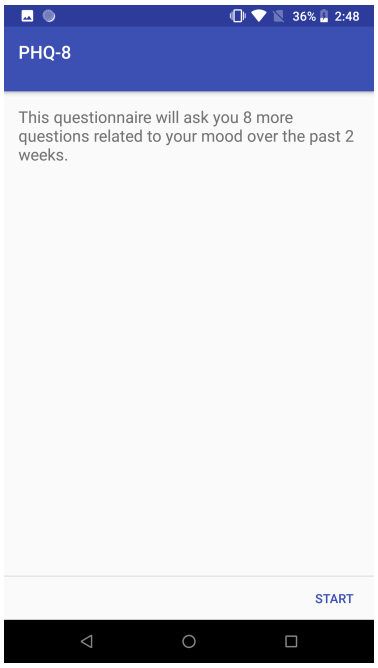
THIS CONCLUDES THE CONSENT GUIDE. ONLY PARTICIPATE IN THIS STUDY IF YOU AGREE WITH ALL THE TERMS STATED ABOVE.

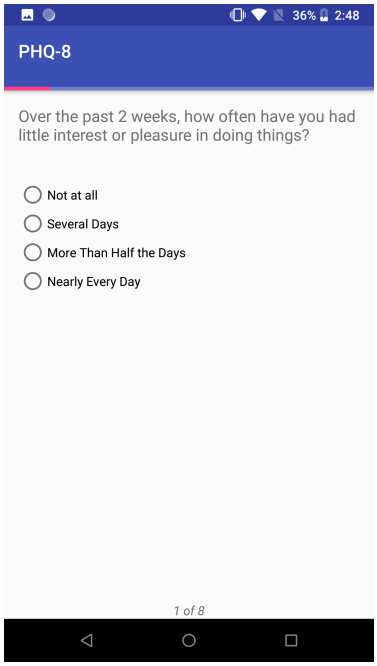
Open study link in a new window

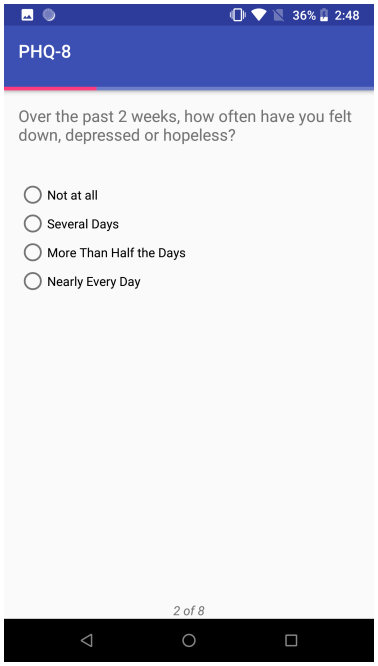
Appendix D

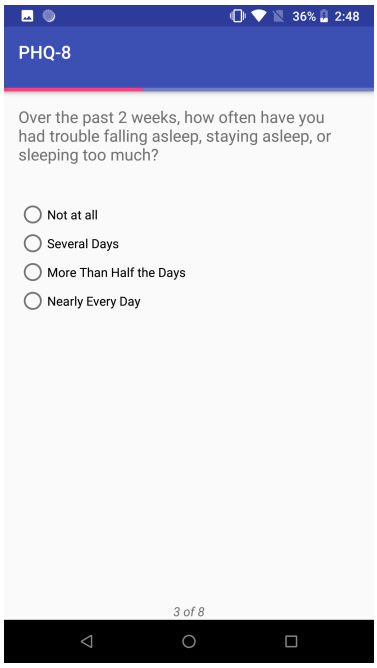
Digitized Self-Report Measures

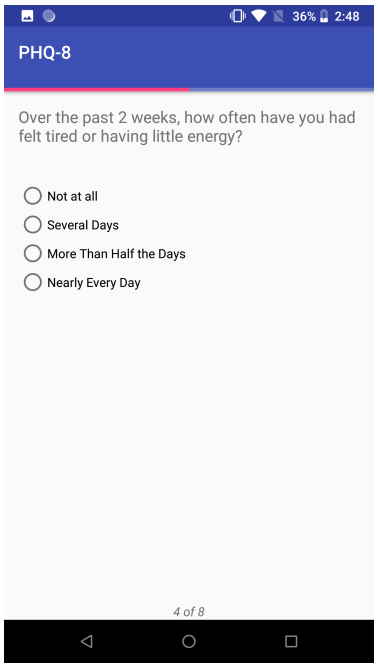

36% 2:46
Questionnaires
You will now be asked to complete 4 questionnaires. Please press START to begin.
START

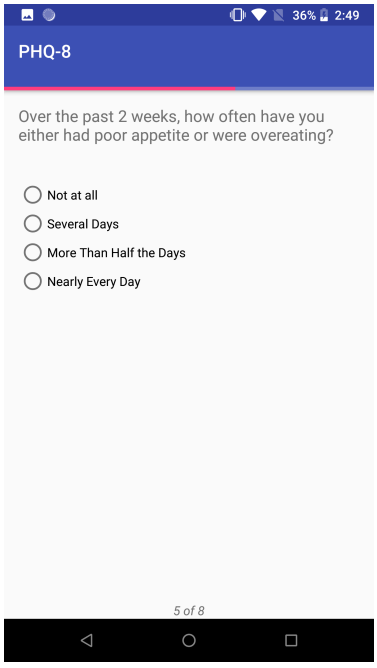

36% 2:48
PHQ-8
This questionnaire will ask you 8 more questions related to your mood over the past 2 weeks.
START

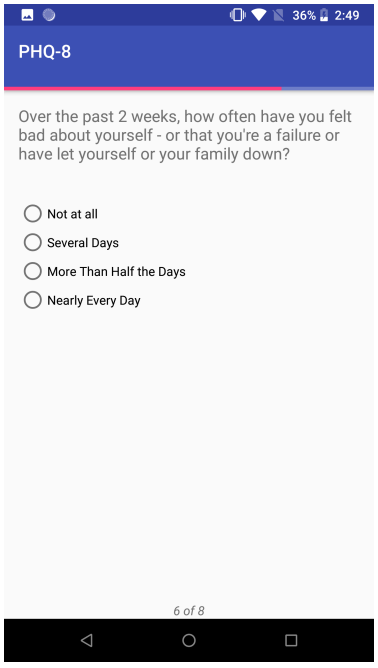

36% 2:48
PHQ-8
Over the past 2 weeks, how often have you had little interest or pleasure in doing things?
Not at all
Several Days
More Than Half the Days
Nearly Every Day
1 of 8

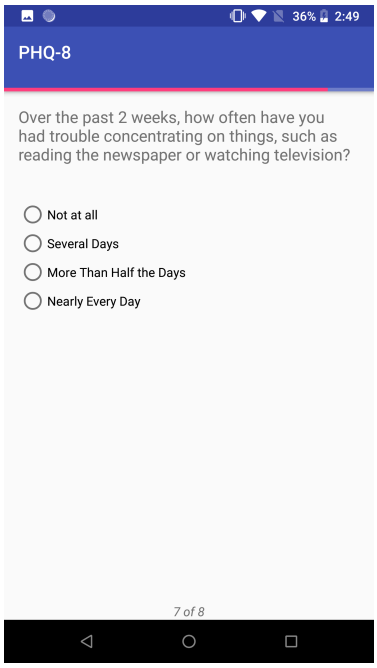

36% 2:48
PHQ-8
Over the past 2 weeks, how often have you felt down, depressed or hopeless?
Not at all
Several Days
More Than Half the Days
Nearly Every Day
2 of 8

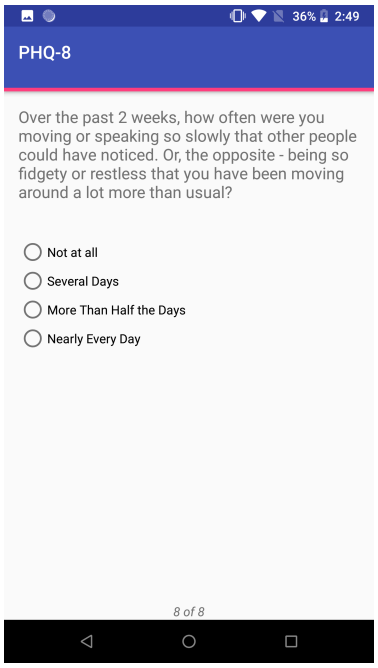

36% 2:48
PHQ-8
Over the past 2 weeks, how often have you had trouble falling asleep, staying asleep, or sleeping too much?
Not at all
Several Days
More Than Half the Days
Nearly Every Day
3 of 8

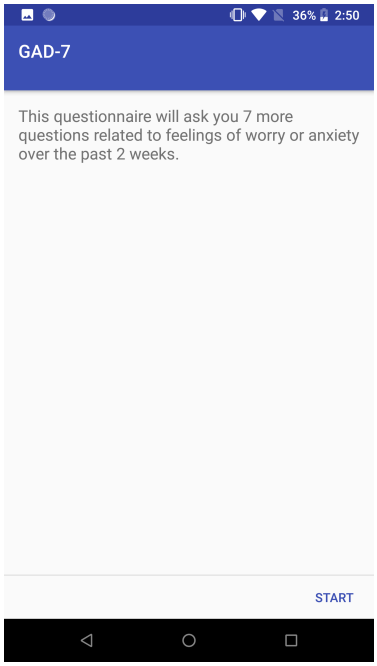

36% 2:48
PHQ-8
Over the past 2 weeks, how often have you had felt tired or having little energy?
Not at all
Several Days
More Than Half the Days
Nearly Every Day
4 of 8

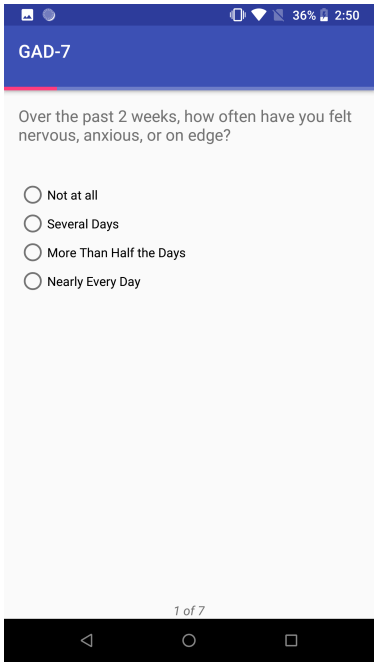

36% 2:49
PHQ-8
Over the past 2 weeks, how often have you either had poor appetite or were overeating?
Not at all
Several Days
More Than Half the Days
Nearly Every Day
5 of 8

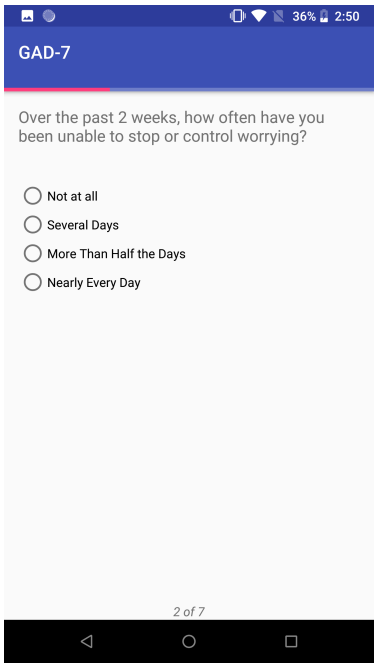

36% 2:49
PHQ-8
Over the past 2 weeks, how often have you felt bad about yourself - or that you're a failure or have let yourself or your family down?
Not at all
Several Days
More Than Half the Days
Nearly Every Day
6 of 8

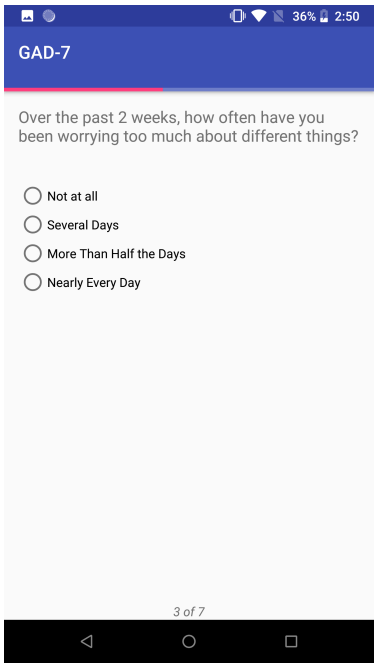

36% 2:49
PHQ-8
Over the past 2 weeks, how often have you had trouble concentrating on things, such as reading the newspaper or watching television?
Not at all
Several Days
More Than Half the Days
Nearly Every Day
7 of 8

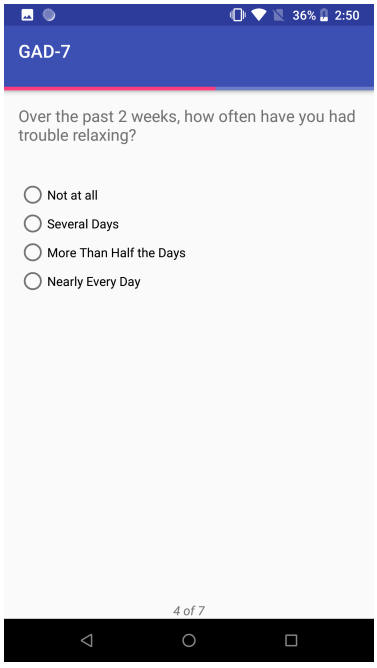

36% 2:49
PHQ-8
Over the past 2 weeks, how often were you moving or speaking so slowly that other people could have noticed. Or, the opposite - being so fidgety or restless that you have been moving around a lot more than usual?
Not at all
Several Days
More Than Half the Days
Nearly Every Day
8 of 8

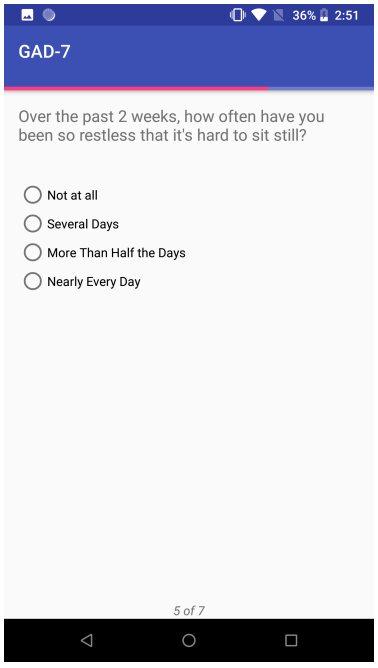

36% 2:50
GAD-7
This questionnaire will ask you 7 more questions related to feelings of worry or anxiety over the past 2 weeks.
START

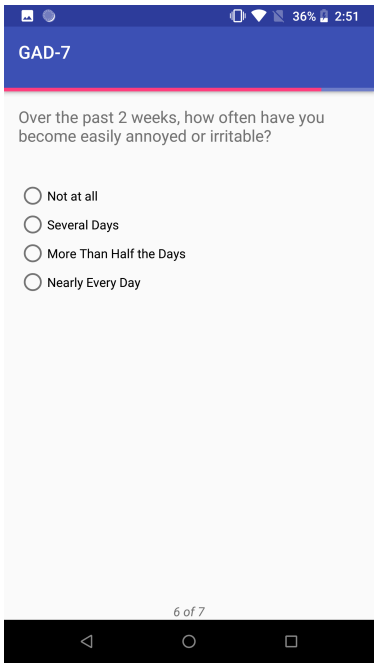

36% 2:50
GAD-7
Over the past 2 weeks, how often have you felt nervous, anxious, or on edge?
Not at all
Several Days
More Than Half the Days
Nearly Every Day
1 of 7

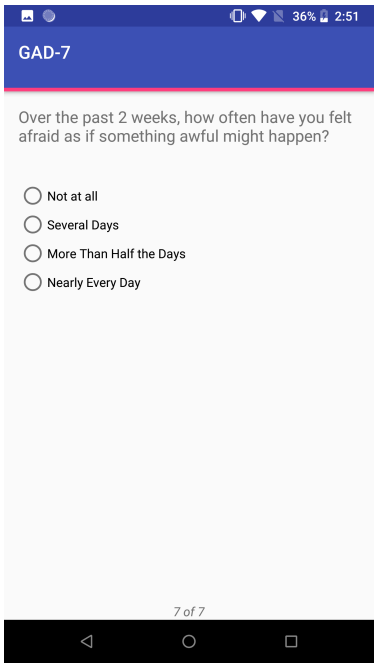

36% 2:50
GAD-7
Over the past 2 weeks, how often have you been unable to stop or control worrying?
Not at all
Several Days
More Than Half the Days
Nearly Every Day
2 of 7


36% 2:50
GAD-7
Over the past 2 weeks, how often have you been worrying too much about different things?
Not at all
Several Days
More Than Half the Days
Nearly Every Day
3 of 7


36% 2:50
GAD-7
Over the past 2 weeks, how often have you had trouble relaxing?
Not at all
Several Days
More Than Half the Days
Nearly Every Day
4 of 7


36% 2:51
GAD-7
Over the past 2 weeks, how often have you been so restless that it's hard to sit still?
Not at all
Several Days
More Than Half the Days
Nearly Every Day
5 of 7


36% 2:51
GAD-7
Over the past 2 weeks, how often have you become easily annoyed or irritable?
Not at all
Several Days
More Than Half the Days
Nearly Every Day
6 of 7


36% 2:51
GAD-7
Over the past 2 weeks, how often have you felt afraid as if something awful might happen?
Not at all
Several Days
More Than Half the Days
Nearly Every Day
7 of 7

Appendix E

Correlations Between Bigram Features and Mental Health Measures

LIWC Category	LSAS		GAD-7		PHQ-8		SDS	
	r	P	r	P	r	P	r	P
achieve	-0.10	0.3784	-0.03	0.8133	-0.08	0.4673	-0.06	0.5638
adj	0.00	0.9734	-0.02	0.8271	-0.03	0.7622	-0.20	0.0678
adverb	0.08	0.4724	0.07	0.5251	0.14	0.2044	0.05	0.6224
affect	-0.15	0.1753	-0.07	0.5419	-0.05	0.6308	-0.10	0.3558
affiliation	0.00	0.9718	-0.02	0.8589	-0.06	0.6096	0.06	0.5606
anger	-0.13	0.2498	0.07	0.5368	0.11	0.2943	-0.01	0.9002
anx	0.04	0.6924	-0.08	0.4742	-0.04	0.7077	0.01	0.9553
article	0.19	0.0867	0.10	0.3683	0.03	0.7647	0.01	0.8943
auxverb	0.07	0.5411	0.04	0.6923	0.12	0.2529	0.05	0.6158
bio	-0.12	0.2641	-0.23	0.0361	-0.15	0.1815	-0.11	0.2952
body	-0.04	0.7041	-0.10	0.3752	-0.04	0.6869	-0.02	0.8525
cause	0.02	0.8416	0.02	0.8203	-0.05	0.6233	-0.07	0.5330
certain	-0.07	0.5120	0.08	0.4538	0.08	0.4507	-0.02	0.8591
cogproc	-0.04	0.7309	0.13	0.2416	0.16	0.1404	0.05	0.6640
compare	0.01	0.9551	0.14	0.2085	0.06	0.5896	-0.03	0.7692
conj	-0.06	0.5860	0.08	0.4436	0.04	0.7411	0.04	0.6974
death	0.32	0.0023	0.27	0.0119	0.41	0.0001	0.28	0.0086
differ	0.02	0.8396	0.09	0.4160	0.17	0.1254	0.09	0.3923
discrep	-0.05	0.6227	0.03	0.8155	0.07	0.5222	0.01	0.9153

Table E.1 continued from previous page

LIWC Category	LSAS		GAD-7		PHQ-8		SDS	
	r	P	r	P	r	P	r	P
drives	-0.16	0.1409	-0.21	0.0540	-0.18	0.0953	-0.07	0.5175
family	0.14	0.2118	0.12	0.2548	0.15	0.1650	0.19	0.0725
feel	-0.04	0.7025	-0.16	0.1433	-0.10	0.3382	-0.10	0.3470
female	0.12	0.2633	0.11	0.3310	0.16	0.1365	0.13	0.2351
focusfuture	-0.15	0.1664	-0.18	0.0989	-0.11	0.3081	0.00	0.9869
focuspast	0.06	0.5608	0.02	0.8869	0.10	0.3752	-0.03	0.8011
focuspresent	-0.01	0.8975	-0.01	0.8961	0.10	0.3400	0.09	0.4069
friend	0.06	0.6038	0.26	0.0159	0.15	0.1703	0.24	0.0263
function	0.07	0.5064	0.18	0.0926	0.24	0.0247	0.15	0.1795
health	0.11	0.3254	-0.18	0.1053	-0.18	0.1068	-0.07	0.4966
hear	0.04	0.6916	0.05	0.6272	0.03	0.8094	0.03	0.7782
home	-0.31	0.0033	-0.12	0.2619	-0.24	0.0279	-0.03	0.7703
i	-0.06	0.5547	0.00	0.9897	0.11	0.3008	0.01	0.9559
ingest	-0.06	0.5701	-0.18	0.0938	-0.13	0.2337	-0.03	0.7824
insight	-0.07	0.5406	0.11	0.3126	0.09	0.4300	0.00	0.9970
interrog	0.01	0.9393	0.03	0.7903	0.05	0.6617	-0.01	0.9578
ipron	0.00	0.9661	0.08	0.4459	0.15	0.1628	0.09	0.4246
leisure	-0.02	0.8668	-0.13	0.2476	-0.09	0.3914	0.02	0.8563
male	0.12	0.2535	0.15	0.1549	0.15	0.1661	0.11	0.2970
money	-0.17	0.1080	-0.14	0.1950	-0.19	0.0779	-0.21	0.0559
motion	0.01	0.8947	-0.05	0.6720	0.07	0.5118	0.19	0.0774
negate	-0.06	0.5900	0.02	0.8843	0.18	0.1001	0.23	0.0325
negemo	-0.11	0.3209	0.09	0.4156	0.15	0.1652	0.04	0.7194
number	0.06	0.5637	-0.03	0.7695	-0.05	0.6178	-0.07	0.4952
percept	0.20	0.0674	0.01	0.9463	0.02	0.8880	0.03	0.8178
posemo	-0.10	0.3656	-0.15	0.1569	-0.18	0.0907	-0.16	0.1460
power	-0.04	0.7110	-0.01	0.9231	0.05	0.6672	0.04	0.7372
ppron	-0.03	0.7936	0.03	0.7514	0.15	0.1803	0.07	0.5257
prep	0.06	0.6022	0.24	0.0293	0.14	0.2077	0.09	0.4010
pronoun	-0.02	0.8640	0.06	0.5708	0.18	0.0924	0.09	0.4061
quant	-0.06	0.5885	0.02	0.8290	-0.08	0.4539	-0.05	0.6624
relativ	0.02	0.8510	-0.22	0.0448	-0.07	0.5223	-0.03	0.7639
relig	0.19	0.0862	0.06	0.5755	0.10	0.3622	0.15	0.1723

Table E.1 continued from previous page

LIWC Category	LSAS		GAD-7		PHQ-8		SDS	
	r	P	r	P	r	P	r	P
reward	-0.17	0.1206	-0.29	0.0069	-0.22	0.0429	-0.10	0.3407
risk	-0.10	0.3510	-0.13	0.2498	-0.08	0.4583	-0.01	0.9624
sad	-0.06	0.6135	0.07	0.5083	0.17	0.1262	0.06	0.5841
see	0.31	0.0034	0.04	0.7467	0.08	0.4848	0.04	0.7299
sexual	-0.24	0.0237	-0.08	0.4915	-0.04	0.7399	-0.16	0.1475
shehe	0.13	0.2446	0.08	0.4587	0.11	0.2973	0.01	0.9109
social	0.02	0.8690	0.12	0.2630	0.11	0.3318	0.14	0.1978
space	0.06	0.5759	-0.14	0.2139	-0.12	0.2756	-0.12	0.2917
tentat	0.03	0.8093	0.05	0.6320	0.14	0.1918	0.06	0.5923
they	0.08	0.4495	0.13	0.2438	0.10	0.3515	0.12	0.2849
time	-0.03	0.8046	-0.20	0.0630	-0.03	0.7497	-0.06	0.6064
verb	0.03	0.7770	0.00	0.9727	0.13	0.2214	0.08	0.4540
we	0.05	0.6477	0.00	0.9723	0.02	0.8447	0.06	0.5862
work	-0.12	0.2823	0.10	0.3373	-0.06	0.5986	-0.08	0.4649
you	-0.11	0.3250	-0.03	0.7572	0.00	0.9910	0.07	0.5444

Table E.1: Correlations between word-usage rates (in each LIWC category) and mental health measures.

Appendix F

R Script for Classification Model Benchmarking

```r
set.seed(1)
library(here)
library("data.table")
library("mlr3")
library("mlr3tuning")
library("mlr3learners")
library("mlr3pipelines")
library("mlr3filters")
library("mlr3viz")
library("mlr3misc")
library("ggplot2")
library("paradox")
library("glmnet")
library("igraph")
library("glmnet")
library("imputeMissings")
library("smotefamily")

#########################
# FUNCTION DEFINITIONS #
#########################
corrected_repeated_kfold_cv_test <- function(a, b, n, k) {
  # 'a' and 'b' are vectors
  if (length(a) != length(a)) {
    stop("unequal number of observations")
  }
  n1 <- k - 1
  n2 <- 1
  diffs <- a - b
  t <- ((1/(k*n)) * sum(diffs)) / sqrt( (1/(k*n) + n2/n1)*var(diffs) )
  # one-tailed pvalue since AUC can only be >= 0.5, not below
  p <- pt(-abs(t), df=k*n-1, lower.tail = TRUE)
  output <- list(t_statistic=t, p_val=p)
}

```

```
##################
# SCRIPT START #
##################

# set up logging
requireNamespace("lgr")
logger = lgr::get_logger("mlr3")
logger$set_threshold("info")
tf = tempfile("mlr3log_", fileext = ".txt")
logger$add_appender(lgr::AppenderFile$new(tf))

# turn on parallelization
future::plan("multiprocess")

# read in data
data = read.csv(here("datasets", "classif_data_some_missing.csv"))

# convert labels to factor data
data$MDD <- as.factor(data$MDD)
data$SAD <- as.factor(data$SAD)
data$GAD <- as.factor(data$GAD)
skimr::skim(data)

# can only do regression with 1 output - drop all scales but one
sad_data = subset(data, select = -c(GAD, MDD))
gad_data = subset(data, select = -c(SAD, MDD))
mdd_data = subset(data, select = -c(GAD, SAD))

# build classification task for each scale
sad_task = TaskClassif$new(id = "SAD", backend = sad_data, target = "SAD", positive =
    "False")
gad_task = TaskClassif$new(id = "GAD", backend = gad_data, target = "GAD", positive =
    "False")
mdd_task = TaskClassif$new(id = "MDD", backend = mdd_data, target = "MDD", positive =
    "False")
my_task_ids = c("SAD", "GAD", "MDD")

# enable stratified sampling:
# First, the observations are divided into subpopulations based on
# the class variable. Then, sampling is done in each of the two subpopulations
# seperately to maintain class balance
sad_task$col_roles$stratum = sad_task$col_roles$target
gad_task$col_roles$stratum = gad_task$col_roles$target
mdd_task$col_roles$stratum = mdd_task$col_roles$target

# resampling and autotuning settings
n_repeats <- 20
k_inner <- 5
k_outer <- 5
max_evals <- 20
metric = msr("classif.logloss")

# MODEL DEFINITIONS

```

```
# 0: baseline model - no features
learner_nofeat = lrn("classif.featureless")
learner_nofeat$predict_type = "prob"
learner_nofeat$id = "Featureless"

# 1: classification tree
feat_scale_tree = po("scale", param_vals = list(center=TRUE, scale=TRUE))
learner_tree = lrn("classif.rpart", predict_type = "prob")
tree_graph = feat_scale_tree %>>% po("imputemedian") %>>% learner_tree
tree_gl = GraphLearner$new(tree_graph)
tree_ps = ParamSet$new(list(
  ParamInt$new("classif.rpart.minsplit", lower = 1, upper = 20)
))
tree_at = AutoTuner$new(
  learner = tree_gl,
  resampling = rsmp("cv", folds = k_inner),
  measure = metric,
  search_space = tree_ps,
  terminator = trm("evals", n_evals = 1*max_evals),
  tuner = tnr("grid_search", resolution = 20),
  store_tuning_instance = FALSE,
  store_benchmark_result = FALSE,
  store_models = FALSE
)
tree_at$id = "DecisionTree"

# 2: Linear Discriminant Analysis
feat_scale_lda = po("scale", param_vals = list(center=TRUE, scale=TRUE))
learner_lda = lrn("classif.lda", predict_type = "prob")
lda_graph = feat_scale_lda %>>% po("imputemedian") %>>% learner_lda
lda_gl = GraphLearner$new(lda_graph)
lda_gl$id = "LDA"

# 3: logistic regression
feat_scale_logreg = po("scale", param_vals = list(center=TRUE, scale=TRUE))
smote_logreg = po("smote", dup_size = 1)
feat_impute1 = po("imputemedian")
learner_logreg = lrn("classif.log_reg", predict_type = "prob")
logreg_graph = feat_scale_logreg %>>% feat_impute1 %>>% learner_logreg
log_reg_gl = GraphLearner$new(logreg_graph)
log_reg_gl$id = "LogisticRegression"

# 4: logistic regression w/ LASSO
enet_learner = lrn("classif.cv_glmnet", predict_type="prob", alpha=1, standardize=
    FALSE)
enet_pv = enet_learner$param_set$values
enet_pv$relax = TRUE
enet_pv$nfolds = 3
enet_learner$param_set$values = enet_pv
enet_learner$param_set
enet_graph = po("imputemedian") %>>% enet_learner
enet_gl = GraphLearner$new(enet_graph)
enet_gl$id = "ElasticNet"

```

```
# 5: naive bayes
feat_scale = po("scale", param_vals = list(center=TRUE, scale=TRUE))
feat_impute1 = po("imputemedian")
learner_nb = lrn("classif.naive_bayes", predict_type = "prob")
nb_graph = feat_scale %>>% po("imputemedian") %>>% learner_nb
naive_bayes_gl = GraphLearner$new(nb_graph)
naive_bayes_gl$id = "NaiveBayes"

# 6: Random Forest - standard trees
feat_scale_forest_fd = po("scale", param_vals = list(center=TRUE, scale=TRUE))
learner_forest_fd = lrn("classif.ranger", importance = "permutation", predict_type =
    "prob",
                        max.depth = 0)
learner_forest_fd$param_set
forest_fd_graph = feat_scale_forest_fd %>>% po("imputemedian") %>>% learner_forest_fd
forest_fd_gl = GraphLearner$new(forest_fd_graph)
forest_fd_ps = ParamSet$new(list(
  ParamInt$new("classif.ranger.num.trees", lower = 10, upper = 1000),
  ParamInt$new("classif.ranger.mtry", lower = 1, upper = 8)
))
forest_fd_at = AutoTuner$new(
  learner = forest_fd_gl,
  resampling = rsmp("cv", folds = k_inner),
  measure = metric,
  search_space = forest_fd_ps,
  terminator = trm("evals", n_evals = 2*max_evals),
  tuner = tnr("random_search"),
  store_tuning_instance = FALSE,
  store_benchmark_result = FALSE,
  store_models = FALSE
)
forest_fd_at$id = "RF_FullDepth"

# 7: Random Forest - stumps
feat_scale_forest_stumps = po("scale", param_vals = list(center=TRUE, scale=TRUE))
learner_forest_stumps = lrn("classif.ranger", importance = "permutation", predict_
    type = "prob",
                        max.depth = 1)
learner_forest_stumps$param_set
forest_stumps_graph = feat_scale_forest_stumps %>>% po("imputemedian") %>>% learner_
    forest_stumps
forest_stumps_gl = GraphLearner$new(forest_stumps_graph)
forest_stumps_ps = ParamSet$new(list(
  ParamInt$new("classif.ranger.num.trees", lower = 10, upper = 500),
  ParamInt$new("classif.ranger.mtry", lower = 1, upper = 4)
))
forest_stumps_at = AutoTuner$new(
  learner = forest_stumps_gl,
  resampling = rsmp("cv", folds = k_inner),
  measure = metric,
  search_space = forest_stumps_ps,
  terminator = trm("evals", n_evals = 2*max_evals),
  tuner = tnr("random_search"),
  store_tuning_instance = FALSE,
```

```
  store_benchmark_result = FALSE,
  store_models = FALSE
)
forest_stumps_at$id = "RF_Stumps"

# 8: linear kernel SVM
feat_scale_svm = po("scale", param_vals = list(center=TRUE, scale=TRUE))
linear_svm = lrn("classif.svm", type = "C-classification",
                 predict_type = "prob", kernel = "linear", id="svm_linear")
linear_svm_graph = feat_scale_svm %>>% po("imputemedian") %>>% linear_svm
linear_svm_gl = GraphLearner$new(linear_svm_graph)
linear_svm_ps = ParamSet$new(list(
  ParamDbl$new("svm_linear.cost", lower = 0.1, upper = 10.0)
))
linear_svm_at = AutoTuner$new(
  learner = linear_svm_gl,
  resampling = rsmp("cv", folds = k_inner),
  measure = metric,
  search_space = linear_svm_ps,
  terminator = trm("evals", n_evals = 1*max_evals),
  tuner = tnr("random_search"),
  store_tuning_instance = FALSE,
  store_benchmark_result = FALSE,
  store_models = FALSE
)
linear_svm_at$id = "SVM_Linear"

# 9: polynomial kernel SVM
feat_scale_svm = po("scale", param_vals = list(center=TRUE, scale=TRUE))
polynomial_svm = lrn("classif.svm", type = "C-classification",
                     predict_type = "prob", kernel = "polynomial", id="svm_polynomial")
polynomial_svm_graph = feat_scale_svm %>>% po("imputemedian") %>>% polynomial_svm
polynomial_svm_gl = GraphLearner$new(polynomial_svm_graph)
polynomial_svm_ps = ParamSet$new(list(
  ParamDbl$new("svm_polynomial.gamma", lower = 0.1, upper = 10.0),
  ParamDbl$new("svm_polynomial.cost", lower = 0.1, upper = 10.0),
  ParamInt$new("svm_polynomial.degree", lower = 2, upper = 3)
))
polynomial_svm_at = AutoTuner$new(
  learner = polynomial_svm_gl,
  resampling = rsmp("cv", folds = k_inner),
  measure = metric,
  search_space = polynomial_svm_ps,
  terminator = trm("evals", n_evals = 3*max_evals),
  tuner = tnr("random_search"),
  store_tuning_instance = FALSE,
  store_benchmark_result = FALSE,
  store_models = FALSE
)
polynomial_svm_at$id = "SVM_Polynomial"

# 10: RBF kernel SVM
feat_scale_svm = po("scale", param_vals = list(center=TRUE, scale=TRUE))
radial_svm = lrn("classif.svm", type = "C-classification",
```

```
                      predict_type = "prob", kernel = "radial", id="svm_radial")
radial_svm_graph = feat_scale_svm %>>% po("imputemedian") %>>% radial_svm
radial_svm_gl = GraphLearner$new(radial_svm_graph)
radial_svm_ps = ParamSet$new(list(
  ParamDbl$new("svm_radial.cost", lower = 0.1, upper = 10.0),
  ParamDbl$new("svm_radial.gamma", lower = 0.1, upper = 10.0)
))
radial_svm_at = AutoTuner$new(
  learner = radial_svm_gl,
  resampling = rsmp("cv", folds = k_inner),
  measure = metric,
  search_space = radial_svm_ps,
  terminator = trm("evals", n_evals = 2*max_evals),
  tuner = tnr("random_search"),
  store_tuning_instance = FALSE,
  store_benchmark_result = FALSE,
  store_models = FALSE
)
radial_svm_at$id = "SVM_RBF"

# 11: sigmoid kernel SVM
feat_scale_svm = po("scale", param_vals = list(center=TRUE, scale=TRUE))
sigmoid_svm = lrn("classif.svm", type = "C-classification",
                  predict_type = "prob", kernel = "sigmoid", id="svm_sigmoid")
sigmoid_svm_graph = feat_scale_svm %>>% po("imputemedian") %>>% sigmoid_svm
sigmoid_svm_gl = GraphLearner$new(sigmoid_svm_graph)
sigmoid_svm_ps = ParamSet$new(list(
  ParamDbl$new("svm_sigmoid.cost", lower = 0.1, upper = 10.0),
  ParamDbl$new("svm_sigmoid.gamma", lower = 0.1, upper = 10.0)
))
sigmoid_svm_at = AutoTuner$new(
  learner = sigmoid_svm_gl,
  resampling = rsmp("cv", folds = k_inner),
  measure = metric,
  search_space = sigmoid_svm_ps,
  terminator = trm("evals", n_evals = 2*max_evals),
  tuner = tnr("random_search"),
  store_tuning_instance = FALSE,
  store_benchmark_result = FALSE,
  store_models = FALSE
)
sigmoid_svm_at$id = "SVM_Sigmoid"

# define train and test using nested resampling
design = benchmark_grid(
  tasks = list(sad_task, gad_task, mdd_task),
  learners = list(learner_nofeat, log_reg_gl, lda_gl, tree_at, forest_fd_at,
    forest_stumps_at, linear_svm_at, sigmoid_svm_at, enet_gl, naive_bayes_gl),
  resamplings = rsmp("repeated_cv", repeats = n_repeats, folds = k_outer)
)

# run the whole set of models
bmr = benchmark(design, store_models = FALSE)

```

```
# get performance measures
measures = msrs(c("classif.ce", "classif.tpr", "classif.tnr",  "classif.auc"))
performances = bmr$aggregate(measures)
perf_table = performances[, c("task_id", "learner_id", "classif.ce",
  "classif.tpr", "classif.tnr", "classif.auc")]

# do my own aggregation of results to get stddev of AUC scores and perform
# t-test to compare to featureless model
print(bmr$score(measures))
all_scores = bmr$score(measures)[, c("task_id", "learner_id", "classif.auc")]
all_scores[, AUC_STDEV := sd(classif.auc), by=list(task_id, learner_id)]
all_scores[, t := corrected_repeated_kfold_cv_test(classif.auc, rep(0.5, n_repeats*k_
    outer), n_repeats, k_outer)[1], by=list(task_id, learner_id)]
all_scores[, p := corrected_repeated_kfold_cv_test(classif.auc, rep(0.5, n_repeats*k_
    outer), n_repeats, k_outer)[2], by=list(task_id, learner_id)]

# join extra measures to existing measures
extra_perf <- unique(all_scores, by=c("task_id", "learner_id"))
extra_perf = extra_perf[,c("task_id", "learner_id", "AUC_STDEV", "t", "p")]
merged = merge(perf_table, extra_perf, by.x=c("task_id", "learner_id"),by.y=c("task_
    id", "learner_id"))

# save model performance scores
fwrite(merged, here("classification_performance.csv"))

# done
warnings()
```

Bibliography

[1] P. Smetanin, D. Stiff, C. Briante, C. E. Adair, S. Ahmad, and M. Khan, "The life and economic impact of major mental illnesses in canada: 2011 to 2041," *RiskAnalytica, on behalf of the Mental Health Commission of Canada*, 2011.

[2] "The impact of mental illness on quality of life: A comparison of severe mental illness, common mental disorder and healthy population samples," *Quality of Life Research*, vol. 16, no. 1, pp. 17–29, 2007. DOI: 10.1007/s11136-006-9002-6. [Online]. Available: https://doi.org/10.1007/s11136-006-9002-6.

[3] E. Chesney, G. M. Goodwin, and S. Fazel, "Risks of all-cause and suicide mortality in mental disorders: a meta-review," *World Psychiatry*, vol. 13, no. 2, pp. 153–160, Jun. 2014.

[4] K. L. Lim, P. Jacobs, A. Ohinmaa, D. Schopflocher, and C. S. Dewa, "A new population-based measure of the economic burden of mental illness in Canada," *Chronic Dis Can*, vol. 28, no. 3, pp. 92–98, 2008.

[5] S. O'Donnell, R. Cheung, K. Bennett, and C. Lagac? "The 2014 Survey on Living with Chronic Diseases in Canada on Mood and Anxiety Disorders: a methodological overview," *Health Promot Chronic Dis Prev Can*, vol. 36, no. 12, pp. 275–288, Dec. 2016.

[6] "Annual report 2016," Office of the Auditor General of Ontario, Tech. Rep., 2016.

[7] D. S. Mennin, D. M. Fresco, R. G. Heimberg, F. R. Schneier, S. O. Davies, and M. R. Liebowitz, "Screening for social anxiety disorder in the clinical setting: Using the liebowitz social anxiety scale," *Journal of Anxiety Disorders*, vol. 16, no. 6, pp. 661–673, 2002, ISSN: 0887-6185. DOI: http://dx.doi.org/10.1016/S0887-6185(02)00134-2. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0887618502001342.

[8] R. L. Spitzer, K. Kroenke, J. B. Williams, and B. Lowe, "A brief measure for assessing generalized anxiety disorder: the GAD-7," *Arch Intern Med*, vol. 166, no. 10, pp. 1092–1097, May 2006.

[9] K. Kroenke and R. L. Spitzer, "The phq-9: A new depression diagnostic and severity measure," *Psychiatric annals*, vol. 32, no. 9, pp. 509–515, 2002.

[10] D. S. Moskowitz and S. N. Young, "Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology," *J Psychiatry Neurosci*, vol. 31, no. 1, pp. 13–20, Jan. 2006.

[11] A. P. Association., *Diagnostic and statistical manual of mental disorders : DSM-5*, English, 5th ed. American Psychiatric Association Arlington, VA, 2013, xliv, 947 p. ISBN: 8123923791.

[12] E. K. Duran, A. W. Aday, N. R. Cook, J. E. Buring, P. M. Ridker, and A. D. Pradhan, "Triglyceride-rich lipoprotein cholesterol, small dense ldl cholesterol, and incident cardiovascular disease," *Journal of the American College of Cardiology*, vol. 75, no. 17, pp. 2122–2135, 2020, ISSN: 0735-1097. DOI: https://doi.org/10.1016/j.jacc.2020.02.059. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0735109720345344.

[13] B. Bandelow, D. Baldwin, M. Abelli, C. Altamura, B. Dell'Osso, K. Domschke, N. A. Fineberg, E. Grünblatt, M. Jarema, E. Maron, *et al.*, "Biological markers for anxiety disorders, ocd and ptsd–a consensus statement. part i: Neuroimaging and genetics," *The World Journal of Biological Psychiatry*, vol. 17, no. 5, pp. 321–365, 2016.

[14] R. Strawbridge, A. H. Young, and A. J. Cleare, "Biomarkers for depression: recent insights, current challenges and future prospects," *Neuropsychiatr Dis Treat*, vol. 13, pp. 1245–1262, 2017.

[15] L. M. Babrak, J. Menetski, M. Rebhan, G. Nisato, M. Zinggeler, N. Brasier, K. Baerenfaller, T. Brenzikofer, L. Baltzer, C. Vogler, L. Gschwind, C. Schneider, F. Streiff, P. M. A. Groenen, and E. Miho, "Traditional and Digital Biomarkers: Two Worlds Apart?" *Digit Biomark*, vol. 3, no. 2, pp. 92–102, 2019. DOI: https://doi.org/10.1159/000502000.

[16] W. H. Organization, *ICD-10 : the ICD-10 classification of mental and behavioural disorders : diagnostic criteria for research.* Geneva, 1993.

[17] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders: DSM-IV*, 4th ed. Washington, DC: Autor, 1994.

[18] G. of Canada, *The Human Face of Mental Health and Mental Illness in Canada.* Minister of Public Works and Government Services Canada, 2006.

[19] R. C. Kessler, K. A. McGonagle, S. Zhao, C. B. Nelson, M. Hughes, S. Eshleman, H.-U. Wittchen, and K. S. Kendler, "Lifetime and 12-Month Prevalence of DSM-III-R Psychiatric Disorders in the United States: Results From the National Comorbidity Survey," *Archives of General Psychiatry*, vol. 51, no. 1, pp. 8–19, Jan. 1994, ISSN: 0003-990X. DOI: 10.1001/archpsyc.1994.03950010008002. [Online]. Available: https://doi.org/10.1001/archpsyc.1994.03950010008002.

[20] L. Marques, A. Chosak, N. M. Simon, D.-M. Phan, S. Wilhelm, and M. Pollack, "Rating scales for anxiety disorders," in *Handbook of Clinical Rating Scales and Assessment in Psychiatry and Mental Health*, L. Baer and M. A. Blais, Eds. Totowa, NJ: Humana Press, 2010, pp. 37–72, ISBN: 978-1-59745-387-5. DOI: 10.1007/978-1-59745-387-5_3. [Online]. Available: http://dx.doi.org/10.1007/978-1-59745-387-5_3.

[21] M. R. Liebowitz, "Social phobia," *Modern Problems in Pharmacopsychiatry*, vol. 22, pp. 141–173, 1987.

[22] S. M. Turner, D. C. Beidel, C. V. Dancu, and M. A. Stanley, "An empirically derived inventory to measure social fears and anxiety: The social phobia and anxiety inventory.," *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, vol. 1.1, pp. 35–40, 1989.

[23] K. M. Connor, J. R. T. Davidson, L. E. Churchill, A. Sherwood, R. H. Weisler, and E. Foa, "Psychometric properties of the social phobia inventory (spin)," *The British Journal of Psychiatry*, vol. 176, no. 4, pp. 379–386, 2000, ISSN: 0007-1250. DOI: 10.1192/bjp.176.4.379. eprint: http://bjp.rcpsych.org/content/176/4/379.full.pdf. [Online]. Available: http://bjp.rcpsych.org/content/176/4/379.

[24] R. P. Mattick and J. Clarke, "Development and validation of measures of social phobia scrutiny fear and social interaction anxiety," *Behaviour Research and Therapy*, vol. 36, no. 4, pp. 455–470, 1998, ISSN: 0005-7967. DOI: http://dx.doi.org/10.1016/S0005-7967(97)10031-6. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0005796797100316.

[25] J. D. Herbert, L. L. Brandsma, and L. Fischer, "Chapter 3 - assessment of social anxiety and its clinical expressions," in *Social Anxiety (Third Edition)*, S. G. Hofmann and P. M. DiBartolo, Eds., Third Edition, San Diego: Academic Press, 2014, pp. 45–94, ISBN: 978-0-12-394427-6. DOI: http://dx.doi.org/10.1016/B978-0-12-394427-6.00003-0. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780123944276000030.

[26] (2021). "Liebowitz Social Anxiety Scale — National Social Anxiety Society," [Online]. Available: https://nationalsocialanxietycenter.com/liebowitz-sa-scale/.

[27] D. M. Fresco, M. E. Coles, R. G. Heimber, M. R. Liebowitz, S. Hami, M. B. Stein, and D. Goetz, "The liebowitz social anxiety scale: A comparison of the psychometric properties of self-report and clinician-administered formats," *Psychological Medicine*, vol. 31, no. 6, pp. 1025–1035, 2001. DOI: 10.1017/S0033291701004056.

[28] K. A. Kobak, S. C. Schaettle, J. H. Greist, J. W. Jefferson, D. J. Katzelnick, and S. L. Dottl, "Computer-administered rating scales for social anxiety in a clinical drug trial," *Depression and Anxiety*, vol. 7, no. 3, pp. 97–104, 1998, ISSN: 1520-6394. DOI: 10.1002/(SICI)1520-6394(1998)7:3<97::AID-DA1>3.0.CO;2-2. [Online]. Available: http://dx.doi.org/10.1002/(SICI)1520-6394(1998)7:3%3C97::AID-DA1%3E3.0.CO;2-2.

[29] N. K. Rytwinski, D. M. Fresco, R. G. Heimberg, M. E. Coles, M. R. Liebowitz, S. Cissell, M. B. Stein, and S. G. Hofmann, "Screening for social anxiety disorder with the self-report version of the liebowitz social anxiety scale," *Depression and Anxiety*, vol. 26, no. 1, pp. 34–38, 2009. DOI: https://doi.org/10.1002/da.20503. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/da.20503. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/da.20503.

[30] R. A. Watterson, J. V. Williams, D. H. Lavorato, and S. B. Patten, "Descriptive Epidemiology of Generalized Anxiety Disorder in Canada," *Can J Psychiatry*, vol. 62, no. 1, pp. 24–29, Jan. 2017.

[31] C. Andreescu, B. H. Belnap, B. L. Rollman, P. Houck, C. Ciliberti, S. Mazumdar, M. K. Shear, and E. J. Lenze, "Generalized anxiety disorder severity scale validation in older adults," *The American Journal of Geriatric Psychiatry*, vol. 16, no. 10, pp. 813–818, 2008, ISSN: 1064-7481. DOI: https://doi.org/10.1097/JGP.0b013e31817c6aab. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1064748112608392.

[32] A. S. Zigmond and R. P. Snaith, "The hospital anxiety and depression scale," *Acta Psychiatr Scand*, vol. 67, no. 6, pp. 361–370, Jun. 1983.

[33] S. B. Patten, J. V. Williams, D. H. Lavorato, J. L. Wang, K. McDonald, and A. G. Bulloch, "Descriptive epidemiology of major depressive disorder in Canada in 2012," *Can J Psychiatry*, vol. 60, no. 1, pp. 23–30, Jan. 2015.

[34] M. Hamilton, "A rating scale for depression," *Journal of neurology, neurosurgery, and psychiatry*, vol. 23, no. 1, p. 56, 1960.

[35] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh, "An inventory for measuring depression," *Archives of general psychiatry*, vol. 4, no. 6, pp. 561–571, 1961.

[36] S. A. Montgomery and M. Åsberg, "A new depression scale designed to be sensitive to change," *British Journal of Psychiatry*, vol. 134, no. 4, pp. 382–389, 1979. DOI: 10.1192/bjp.134.4.382.

[37] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The phq-8 as a measure of current depression in the general population," *Journal of affective disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[38] A. C. Leon, M. Olfson, L. Portera, L. Farber, and D. V. Sheehan, "Assessing psychiatric impairment in primary care with the sheehan disability scale," *The international journal of psychiatry in medicine*, vol. 27, no. 2, pp. 93–105, 1997.

[39] (2021). "Sheehan Disability Scale (SDS) - Harm Research," [Online]. Available: https://harmresearch.org/index.php/product/sheehan-disability-scale-sds-2/.

[40] S. Graham, C. Depp, E. E. Lee, C. Nebeker, X. Tu, H. C. Kim, and D. V. Jeste, "Artificial Intelligence for Mental Health and Mental Illnesses: an Overview," *Curr Psychiatry Rep*, vol. 21, no. 11, p. 116, Nov. 2019.

[41] R. M. Benca, M. Okawa, M. Uchiyama, S. Ozaki, T. Nakajima, K. Shibui, and W. H. Obermeyer, "Sleep and mood disorders," *Sleep Medicine Reviews*, vol. 1, no. 1, pp. 45–56, Nov. 1997.

[42] J. Bueno-Antequera and D. Munguía-Izquierdo, "Exercise and Depressive Disorder," *Adv Exp Med Biol*, vol. 1228, pp. 271–287, 2020.

[43] Z. Chen, M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury, and A. T. Campbell, "Unobtrusive sleep monitoring using smartphones," in *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, IEEE, 2013, pp. 145–152.

[44] S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, and D. C. Mohr, "Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study," *J Med Internet Res*, vol. 17, no. 7, e175, Jul. 2015.

[45] Demasi, Aguilera, and Recht, "Detecting change in depressive symptoms from daily wellbeing questions, personality, and activity," in *2016 IEEE Wireless Health (WH)*, 2016, pp. 1–8. DOI: 10.1109/WH.2016.7764552. [Online]. Available: https://ieeexplore.ieee.org/document/7764552.

[46] D. C. Mohr, M. Zhang, and S. M. Schueller, "Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning," *Annu Rev Clin Psychol*, vol. 13, pp. 23–47, May 2017.

[47] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, 2014, pp. 3–14. DOI: https://doi.org/10.1145/2632048.2632054. [Online]. Available: https://dl.acm.org/doi/10.1145/2632048.2632054.

[48] D. Ben-Zeev, E. A. Scherer, R. Wang, H. Xie, and A. T. Campbell, "Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health," *Psychiatr Rehabil J*, vol. 38, no. 3, pp. 218–226, Sep. 2015. DOI: https://doi.org/10.1037/prj0000130. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25844912/.

[49] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," *J Health Soc Behav*, vol. 24, no. 4, pp. 385–396, Dec. 1983.

[50] P. I. Chow, K. Fua, Y. Huang, W. Bonelli, H. Xiong, L. E. Barnes, and B. A. Teachman, "Using Mobile Sensing to Test Clinical Models of Depression, Social Anxiety, State Affect, and Social Isolation Among College Students," *J Med Internet Res*, vol. 19, no. 3, e62, Mar. 2017.

[51] R. P. Mattick and J. C. Clarke, "Development and validation of measures of social phobia scrutiny fear and social interaction anxiety," *Behav Res Ther*, vol. 36, no. 4, pp. 455–470, Apr. 1998.

[52] M. Boukhechba, P. Chow, K. Fua, B. A. Teachman, and L. E. Barnes, "Predicting Social Anxiety From Global Positioning System Traces of College Students: Feasibility Study," *JMIR Ment Health*, vol. 5, no. 3, e10101, Jul. 2018.

[53] R. Wang, W. Wang, A. DaSilva, J. F. Huckins, W. M. Kelley, T. F. Heatherton, and A. T. Campbell, "Tracking depression dynamics in college students using mobile phone and wearable sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–26, 2018.

[54] L. Canzian and M. Musolesi, "Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015, pp. 1293–1304.

[55] S. Saeb, E. G. Lattie, S. M. Schueller, K. P. Kording, and D. C. Mohr, "The relationship between mobile phone location sensor data and depressive symptom severity," *PeerJ*, vol. 4, e2537, 2016.

[56] A. A. Farhan, C. Yue, R. Morillo, S. Ware, J. Lu, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang, "Behavior vs. introspection: Refining prediction of clinical depression via smartphone sensing data," in *2016 IEEE Wireless Health (WH)*, IEEE, 2016, pp. 1–8.

[57] A. Pratap, D. C. Atkins, B. N. Renn, M. J. Tanana, S. D. Mooney, J. A. Anguera, and P. A. Areán, "The accuracy of passive phone sensors in predicting daily mood," *Depression and anxiety*, vol. 36, no. 1, pp. 72–81, 2019.

[58] P. Chikersal, A. Doryab, M. Tumminia, D. K. Villalba, J. M. Dutcher, X. Liu, S. Cohen, K. G. Creswell, J. Mankoff, J. D. Creswell, *et al.*, "Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: A machine learning approach with robust feature selection," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 28, no. 1, pp. 1–41, 2021.

[59] S. A. Thompson and C. Warzel, "Twelve million phones, one dataset, zero privacy," *The New York Times*, Dec. 2019. [Online]. Available: https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html.

[60] C. Nebeker, K. Murray, C. Holub, J. Haughton, and E. M. Arredondo, "Acceptance of Mobile Health in Communities Underrepresented in Biomedical Research: Barriers and Ethical Considerations for Scientists," *JMIR Mhealth Uhealth*, vol. 5, no. 6, e87, Jun. 2017.

[61] S. Sicari, A. Rizzardi, L. Grieco, and A. Coen-Porisini, "Security, privacy and trust in internet of things: The road ahead," *Computer Networks*, vol. 76, pp. 146–164, 2015, ISSN: 1389-1286. DOI: https://doi.org/10.1016/j.comnet.2014.11.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128614003971.

[62] R. H. Weber, "Internet of things: Privacy issues revisited," *Computer Law & Security Review*, vol. 31, no. 5, pp. 618–627, 2015, ISSN: 0267-3649. DOI: https://doi.org/10.1016/j.clsr.2015.07.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0267364915001156.

[63] B. D. Weinberg, G. R. Milne, Y. G. Andonova, and F. M. Hajjat, "Internet of things: Convenience vs. privacy and secrecy," *Business Horizons*, vol. 58, no. 6, pp. 615–624, 2015, SPECIAL ISSUE: THE MAGIC OF SECRETS, ISSN: 0007-6813. DOI: https://doi.org/10.1016/j.bushor.2015.06.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0007681315000865.

[64] Y. Gao, H. Li, and Y. Luo, "An empirical study of wearable technology acceptance in healthcare," *Industrial Management & Data Systems*, vol. 115, no. 9, pp. 1704–1723, 2015. DOI: https://doi.org/10.1108/IMDS-03-2015-0087.

[65] H. Li, J. Wu, Y. Gao, and Y. Shi, "Examining individuals' adoption of healthcare wearable devices: An empirical study from privacy calculus perspective," *Int J Med Inform*, vol. 88, pp. 8–17, Apr. 2016.

[66] J. Torous, R. Friedman, and M. Keshavan, "Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions," *JMIR Mhealth Uhealth*, vol. 2, no. 1, e2, Jan. 2014.

[67] *Number of smartphone users in the united states from 2018 to 2024 (in millions)*, Apr. 2020. [Online]. Available: https://www.statista.com/statistics/201182/forecast-of-smartphone-users-in-the-us/.

[68] *Life in the fast lane: How are canadians managing?, 2016*, Nov. 2017. [Online]. Available: https://www150.statcan.gc.ca/n1/en/daily-quotidien/171114/dq171114a-eng.pdf.

[69] *Smartphone sales by year by os 2009-2018—statista*, Aug. 2018. [Online]. Available: https://www.statista.com/statistics/266219/global-smartphone-sales-since-1st-quarter-2009-by-operating-system/.

[70] I. McGraw, R. Prabhavalkar, R. Alvarez, M. G. Arenas, K. Rao, D. Rybach, O. Alsharif, H. Sak, A. Gruenstein, F. Beaufays, and C. Parada, "Personalized speech recognition on mobile devices," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5955–5959. DOI: 10.1109/ICASSP.2016.7472820.

[71] P. Andriotis, A. Takasu, and T. Tryfonas, "Smartphone message sentiment analysis," in *Advances in Digital Forensics X*, G. Peterson and S. Shenoi, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 253–265, ISBN: 978-3-662-44952-3.

[72] C. K. Rubanovich, D. C. Mohr, and S. M. Schueller, "Health App Use Among Individuals With Symptoms of Depression and Anxiety: A Survey Study With Thematic Coding," *JMIR Ment Health*, vol. 4, no. 2, e22, Jun. 2017.

[73] J. Torous, S. R. Chan, S. Yee-Marie Tan, J. Behrens, I. Mathew, E. J. Conrad, L. Hinton, P. Yellowlees, and M. Keshavan, "Patient Smartphone Ownership and Interest in Mobile Apps to Monitor Symptoms of Mental Health Conditions: A Survey in Four Geographically Distinct Psychiatric Clinics," *JMIR Ment Health*, vol. 1, no. 1, e5, 2014.

[74] H. Zhang, H. Zhang, X. Wang, Z. Yang, and Y. Zhao, "Analysis of Requirements for Developing an mHealth-Based Health Management Platform," *JMIR Mhealth Uhealth*, vol. 5, no. 8, e117, Aug. 2017.

[75] A. P. Felt, S. Egelman, and D. Wagner, "I've got 99 problems, but vibration ain't one: A survey of smartphone users' concerns," in *Proceedings of the Second ACM Workshop on Security and Privacy in Smartphones and Mobile Devices*, ser. SPSM '12, Raleigh, North Carolina, USA: Association for Computing Machinery, 2012, pp. 33–44, ISBN: 9781450316668. DOI: 10.1145/2381934.2381943. [Online]. Available: https://doi.org/10.1145/2381934.2381943.

[76] E. Fife and J. Orjuela, "The privacy calculus: Mobile apps and user perceptions of privacy and security," *International Journal of Engineering Business Management*, vol. 4, p. 11, 2012. DOI: 10.5772/51645. eprint: https://doi.org/10.5772/51645. [Online]. Available: https://doi.org/10.5772/51645.

[77] Prolific. (2021). "Prolific," [Online]. Available: https://prolific.co/.

[78] Amazon. (2021). "Amazon Mechanical Turk," [Online]. Available: https://www.mturk.com/.

[79] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti, "Beyond the turk: Alternative platforms for crowdsourcing behavioral research," *Journal of Experimental Social Psychology*, vol. 70, pp. 153–163, 2017, ISSN: 0022-1031. DOI: https://doi.org/10.1016/j.jesp.2017.01.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022103116303201.

[80] J. S. Marcano Belisario, J. Jamsek, K. Huckvale, J. O'Donoghue, C. P. Morrison, and J. Car, "Comparison of self-administered survey questionnaire responses collected using mobile apps versus other methods," *Cochrane Database Syst Rev*, no. 7, MR000042, Jul. 2015. DOI: https://doi.org/10.1002/14651858.mr000042.pub2.

[81] (2021). "Criminal Code (R.S.C., 1985, c. C-46)," [Online]. Available: https://laws-lois.justice.gc.ca/eng/acts/c-46/section-184.html.

[82] Google. (2021). "Logger — Apps on Google Play," [Online]. Available: https://play.google.com/store/apps/details?id=ca.utoronto.ece.cimsah.logger (visited on 03/12/2021).

[83] Apple. (2021). "Enabling Background Sessions — Apple Developer Documentation," [Online]. Available: https://developer.apple.com/documentation/watchkit/background_execution/enabling_background_sessions (visited on 03/12/2021).

[84] Google. (2021). "Google Awareness API — Google Developers," [Online]. Available: https://developers.google.com/awareness (visited on 03/10/2021).

[85] ——, (2021). "Sensor — Android Developers," [Online]. Available: https://developer.android.com/reference/android/hardware/Sensor (visited on 03/10/2021).

[86] ——, (2021). "MediaRecorder — Android Developers," [Online]. Available: https://developer.android.com/reference/android/media/MediaRecorder (visited on 03/10/2021).

[87] ——, (2021). "MediaRecorder.AudioEncoder — Android Developers," [Online]. Available: https://developer.android.com/reference/android/media/MediaRecorder.AudioEncoder (visited on 03/10/2021).

[88] ——, (2021). "MediaRecorder.OutputFormat — Android Developers," [Online]. Available: https://developer.android.com/reference/android/media/MediaRecorder.OutputFormat (visited on 03/10/2021).

[89] ——, (2021). "ContactsContract — Android Developers," [Online]. Available: https://developer.android.com/reference/android/provider/ContactsContract (visited on 03/12/2021).

[90] ——, (2021). "CalendarContract — Android Developers," [Online]. Available: https://developer.android.com/reference/android/provider/CalendarContract (visited on 03/12/2021).

[91] ——, (2021). "BroadcastReceiver — Android Developers," [Online]. Available: https://developer.android.com/reference/android/content/BroadcastReceiver (visited on 03/12/2021).

[92] ——, (2021). "Intent — Android Developers," [Online]. Available: https://developer.android.com/reference/android/content/Intent#ACTION_SCREEN_ON (visited on 03/12/2021).

[93] W. Stallings, *Cryptography and network security : principles and practice*. Boston: Pearson, 2017, ISBN: 9780134444284.

[94] Google. (2021). "Javax.crypto — Android Developers," [Online]. Available: https://developer.android.com/reference/javax/crypto/package-summary (visited on 03/13/2021).

[95] ——, (2021). "Firebase," [Online]. Available: https://firebase.google.com/ (visited on 03/13/2021).

[96] ——, (2021). "Firebase Hosting," [Online]. Available: https://firebase.google.com/docs/hosting (visited on 03/13/2021).

[97] ——, (2021). "Cloud Functions for Firebase," [Online]. Available: https://firebase.google.com/docs/functions (visited on 03/13/2021).

[98] ——, (2021). "Firebase Authentication," [Online]. Available: https://firebase.google.com/docs/auth (visited on 03/13/2021).

[99] ——, (2021). "Cloud Firestore — Firebase," [Online]. Available: https://firebase.google.com/docs/auth (visited on 03/13/2021).

[100] ——, (2021). "com.google.firebase.firestore — Firebase," [Online]. Available: https://firebase.google.com/docs/reference/android/com/google/firebase/firestore/package-summary (visited on 03/13/2021).

[101] ——, (2021). "Cloud Storage for Firebase — Firebase," [Online]. Available: https://firebase.google.com/docs/storage (visited on 03/13/2021).

[102] FFmpeg Developers. (2021). "FFmpeg," [Online]. Available: https://www.ffmpeg.org/ (visited on 03/13/2021).

[103] Google. (2021). "Speech-to-Text: Automatic Speech Recognition — Google Cloud," [Online]. Available: https://cloud.google.com/speech-to-text (visited on 03/13/2021).

[104] ——, (2021). "Add the Firebase Admin SDK to your server," [Online]. Available: https://firebase.google.com/docs/admin/setup (visited on 03/13/2021).

[105] ——, (2021). "Optimize for Doze and App Standy— Android Developers," [Online]. Available: https://developer.android.com/training/monitoring-device-state/doze-standby (visited on 03/28/2021).

[106] ——, (2021). "App Standy Buckets— Android Developers," [Online]. Available: https://developer.android.com/topic/performance/appstandby (visited on 03/28/2021).

[107] ——, (2020). "Introducing Amazon Halo and Amazon Halo Band," [Online]. Available: https://www.businesswire.com/news/home/20200827005453/en/ (visited on 03/28/2021).

[108] S. Badillo, B. Banfai, F. Birzele, I. I. Davydov, L. Hutchinson, T. Kam-Thong, J. Siebourg-Polster, B. Steiert, and J. D. Zhang, "An Introduction to Machine Learning," *Clin Pharmacol Ther*, vol. 107, no. 4, pp. 871–885, Apr. 2020.

[109] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96, Portland, Oregon: AAAI Press, 1996, pp. 226–231.

[110] G. Contributors. (2021). "Calculating Distance," [Online]. Available: https://geopy.readthedocs.io/en/stable/#module-geopy.distance (visited on 04/23/2021).

[111] J. T. VanderPlas, "Understanding the lomb–scargle periodogram," *The Astrophysical Journal Supplement Series*, vol. 236, no. 1, p. 16, May 2018. DOI: 10.3847/1538-4365/aab766. [Online]. Available: https://doi.org/10.3847/1538-4365/aab766.

[112] "Fourier transforms and fourier integrals," in *Fourier Analysis*. John Wiley & Sons, Ltd, 2005, ch. 6, pp. 297–352, ISBN: 9781118165508. DOI: https://doi.org/10.1002/9781118165508.ch6. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118165508.ch6. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118165508.ch6.

[113] R. N. Carleton, "Into the unknown: A review and synthesis of contemporary models involving uncertainty," *J Anxiety Disord*, vol. 39, pp. 30–43, Apr. 2016.

[114] E. L. Gentes and A. M. Ruscio, "A meta-analysis of the relation of intolerance of uncertainty to symptoms of generalized anxiety disorder, major depressive disorder, and obsessive-compulsive disorder," *Clin Psychol Rev*, vol. 31, no. 6, pp. 923–933, Aug. 2011.

[115] S. Thomée, A. Härenstam, and M. Hagberg, "Mobile phone use and stress, sleep disturbances, and symptoms of depression among young adults–a prospective cohort study," *BMC Public Health*, vol. 11, p. 66, Jan. 2011.

[116] Y.-K. Lee, C.-T. Chang, Y. Lin, and Z.-H. Cheng, "The dark side of smartphone usage: Psychological traits, compulsive behavior and technostress," *Computers in Human Behavior*, vol. 31, pp. 373–383, 2014, ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2013.10.047. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S074756321300397X.

[117] S. M. Sabet, N. D. Dautovich, and J. M. Dzierzewski, "The Rhythm is Gonna Get You: Social Rhythms, Sleep, Depressive, and Anxiety Symptoms," *J Affect Disord*, vol. 286, pp. 197–203, May 2021. DOI: https://doi.org/10.1016/j.jad.2021.02.061.

[118] M. K. Shear, J. Randall, T. H. Monk, A. Ritenour, X. Tu, E. Frank, C. Reynolds, and D. J. Kupfer, "Social rhythm in anxiety disorder patients," *Anxiety*, vol. 1, no. 2, pp. 90–95, 1994. DOI: https://doi.org/10.1002/anxi.3070010208. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/9160553/.

[119] M. P. Szuba, A. Yager, B. H. Guze, E. M. Allen, and L. R. Baxter, "Disruption of social circadian rhythms in major depression: a preliminary report," *Psychiatry Res*, vol. 42, no. 3, pp. 221–230, Jun. 1992. DOI: https://doi.org/10.1016/0165-1781(92)90114-i. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/1496054/.

[120] R. M. Benca, M. Okawa, M. Uchiyama, S. Ozaki, T. Nakajima, K. Shibui, and W. H. Obermeyer, "Sleep and mood disorders," *Sleep Med Rev*, vol. 1, no. 1, pp. 45–56, Nov. 1997.

[121] D. Neckelmann, A. Mykletun, and A. A. Dahl, "Chronic insomnia as a risk factor for developing anxiety and depression," *Sleep*, vol. 30, no. 7, pp. 873–880, Jul. 2007.

[122] J. A. Gubner, *Probability and Random Processes for Electrical and Computer Engineers*. USA: Cambridge University Press, 2006, ISBN: 0521864704.

[123] H. Atalay, "Comorbidity of insomnia detected by the Pittsburgh sleep quality index with anxiety, depression and personality disorders," *Isr J Psychiatry Relat Sci*, vol. 48, no. 1, pp. 54–59, 2011. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21572244/.

[124] H. J. Ramsawh, M. B. Stein, S. L. Belik, F. Jacobi, and J. Sareen, "Relationship of anxiety disorders, sleep quality, and functional impairment in a community sample," *J Psychiatr Res*, vol. 43, no. 10, pp. 926–933, Jul. 2009. DOI: https://doi.org/10.1016/j.jpsychires.2009.01.009. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/19269650/.

[125] Y. Hayashino, S. Yamazaki, M. Takegami, T. Nakayama, S. Sokejima, and S. Fukuhara, "Association between number of comorbid conditions, depression, and sleep quality using the Pittsburgh Sleep Quality Index: results from a population-based survey," *Sleep Med*, vol. 11, no. 4, pp. 366–371, Apr. 2010. DOI: https://doi.org/10.1016/j.sleep.2009.05.021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20219425/.

[126] M. Y. Ağargün, H. Kara, and M. Solmaz, "Subjective sleep quality and suicidality in patients with major depression," *J Psychiatr Res*, vol. 31, no. 3, pp. 377–381, 1997. DOI: https://doi.org/10.1016/s0022-3956(96)00037-4. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/9306295/.

[127] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin, 2015.

[128] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010. DOI: 10.1177/0261927X09351676. [Online]. Available: https://doi.org/10.1177/0261927X09351676.

[129] J. M. Bland and D. G. Altman, "Multiple significance tests: The bonferroni method," *BMJ*, vol. 310, no. 6973, p. 170, 1995, ISSN: 0959-8138. DOI: 10.1136/bmj.310.6973.170. eprint: https://www.bmj.com/content/310/6973/170.full.pdf. [Online]. Available: https://www.bmj.com/content/310/6973/170.

[130] B. R. Veltman, "Linguistic analysis of the semantic content of the rorschach inkblot test," Ph.D. dissertation, 2005. [Online]. Available: https://search.proquest.com/docview/304953299.

[131] S. W. Stirman and J. W. Pennebaker, "Word use in the poetry of suicidal and nonsuicidal poets," *Psychosom Med*, vol. 63, no. 4, pp. 517–522, 2001.

[132] T. V. Perneger, "What's wrong with Bonferroni adjustments," *BMJ*, vol. 316, no. 7139, pp. 1236–1238, Apr. 1998.

[133] L. M. McTeague, J. R. Shumen, M. J. Wieser, P. J. Lang, and A. Keil, "Social vision: sustained perceptual enhancement of affective facial cues in social anxiety," *Neuroimage*, vol. 54, no. 2, pp. 1615–1624, Jan. 2011.

[134] J. A. Cooper, A. R. Arulpragasam, and M. T. Treadway, "Anhedonia in depression: biological mechanisms and computational models," *Curr Opin Behav Sci*, vol. 22, pp. 128–135, Aug. 2018.

[135] T. Sternat and M. A. Katzman, "Neurobiology of hedonic tone: the relationship between treatment-resistant depression, attention-deficit hyperactivity disorder, and substance abuse," *Neuropsychiatr Dis Treat*, vol. 12, pp. 2149–2164, 2016.

[136] J. A. Gray and N. McNaughton, *The neuropsychology of anxiety : an enquiry into the functions of the septo-hippocampal system*. Oxford New York: Oxford University Press, 2000, ISBN: 9780198522713.

[137] D. Liaqat, R. Wu, A. Gershon, H. Alshaer, F. Rudzicz, and E. de Lara, "Challenges with real-world smartwatch based audio monitoring," in *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*, ser. WearSys '18, Munich, Germany: Association for Computing Machinery, 2018, pp. 54–59, ISBN: 9781450358422. DOI: 10.1145/3211960.3211977. [Online]. Available: https://doi.org/10.1145/3211960.3211977.

[138] T. Joiner and J. Katz, "Contagion of depressive symptoms and mood: Meta-analytic review and explanations from cognitive, behavioral, and interpersonal viewpoints.," *Clinical Psychology: Science and Practice*, vol. 6, pp. 149–164, 1999. DOI: https://doi.apa.org/doi/10.1093/clipsy.6.2.149.

[139] B. Till, U. S. Tran, M. Voracek, G. Sonneck, and T. Niederkrotenthaler, "Associations between film preferences and risk factors for suicide: an online survey," *PLoS One*, vol. 9, no. 7, e102293, 2014.

[140] T. Edwards and N. S. Holtzman, "A meta-analysis of correlations between depression and first person singular pronoun use," *Journal of Research in Personality*, vol. 68, pp. 63–68, 2017, ISSN: 0092-6566. DOI: https://doi.org/10.1016/j.jrp.2017.02.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092656616302884.

[141] M. Settanni and D. Marengo, "Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts," *Front Psychol*, vol. 6, p. 1045, 2015.

[142] J. C. Eichstaedt, R. J. Smith, R. M. Merchant, L. H. Ungar, P. Crutchley, D. Preoţiuc-Pietro, D. A. Asch, and H. A. Schwartz, "Facebook language predicts depression in medical records," *Proc Natl Acad Sci U S A*, vol. 115, no. 44, pp. 11 203–11 208, Oct. 2018.

[143] A. S. Miner, A. Haque, J. A. Fries, S. L. Fleming, D. E. Wilfley, G. Terence Wilson, A. Milstein, D. Jurafsky, B. A. Arnow, W. Stewart Agras, L. Fei-Fei, and N. H. Shah, "Assessing the accuracy of automatic speech recognition for psychotherapy," *NPJ Digit Med*, vol. 3, p. 82, 2020.

[144] M. A. Babyak, "What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models," *Psychosomatic medicine*, vol. 66, no. 3, pp. 411–421, 2004. [Online]. Available: https://journals.lww.com/psychosomaticmedicine/Fulltext/2004/05000/What_You_See_May_Not_Be_What_You_Get__A_Brief.21.

[145] M. Lang, M. Binder, J. Richter, P. Schratz, F. Pfisterer, S. Coors, Q. Au, G. Casalicchio, L. Kotthoff, and B. Bischl, "mlr3: A modern object-oriented machine learning framework in R," *Journal of Open Source Software*, Dec. 2019. DOI: 10.21105/joss.01903. [Online]. Available: https://joss.theoj.org/papers/10.21105/joss.01903.

[146] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*. Boca Raton: Routledge, 2017, ISBN: 9781315139470. DOI: https://doi.org/10.1201/9781315139470.

[147] M. Lang, B. Bischl, J. Richter, P. Schratz, and M. Binder. (2021). "Classification Tree Learner," [Online]. Available: https://mlr3.mlr-org.com/reference/mlr_learners_classif.rpart.html (visited on 05/09/2021).

[148] W. Venables and B. Ripley, *Modern Applied Statistics with S*. New York: Springer-Verlag, 2002, ISBN: 978-0-387-21706-2. DOI: https://doi.org/10.1007/978-0-387-21706-2.

[149] M. Lang, B. Bischl, J. Richter, P. Schratz, and M. Binder. (2021). "Linear Discriminant Analysis Classification Learner," [Online]. Available: https://mlr3learners.mlr-org.com/reference/mlr_learners_classif.lda.html (visited on 05/09/2021).

[150] J. C. Stoltzfus, "Logistic regression: a brief primer," *Acad Emerg Med*, vol. 18, no. 10, pp. 1099–1104, Oct. 2011. DOI: https://doi.org/10.1111/j.1553-2712.2011.01185.x. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21996075/.

[151] M. Lang, B. Bischl, J. Richter, P. Schratz, and M. Binder. (2021). "Logistic Regression Classification Learner," [Online]. Available: https://mlr3learners.mlr-org.com/reference/mlr_learners_classif.log_reg.html (visited on 05/09/2021).

[152] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[153] M. Lang, B. Bischl, J. Richter, P. Schratz, and M. Binder. (2021). "GLM with Elastic Net Regularization Classification Learner," [Online]. Available: https://mlr3learners.mlr-org.com/reference/mlr_learners_classif.glmnet.html (visited on 05/09/2021).

[154] J. Friedman, T. Hastie, R. Tibshirani, *et al.*, *The Elements of Statistical Learning*, 10. Springer series in statistics New York, 2001, vol. 1.

[155] M. Lang, B. Bischl, J. Richter, P. Schratz, and M. Binder. (2021). "Naive Bayes Classification Learner," [Online]. Available: https://mlr3learners.mlr-org.com/reference/mlr_learners_classif.naive_bayes.html (visited on 05/09/2021).

[156] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[157] M. Lang, B. Bischl, J. Richter, P. Schratz, and M. Binder. (2021). "Ranger Classification Learner," [Online]. Available: https://mlr3learners.mlr-org.com/reference/mlr_learners_classif.ranger.html (visited on 05/09/2021).

[158] ——, (2021). "Support Vector Machine," [Online]. Available: https://mlr3learners.mlr-org.com/reference/mlr_learners_classif.svm.html (visited on 05/09/2021).

[159] B. Bischl, O. Mersmann, H. Trautmann, and C. Weihs, "Resampling methods for meta-model validation with recommendations for evolutionary computation," *Evolutionary computation*, vol. 20, no. 2, pp. 249–275, 2012.

[160] M. Becker, M. Binder, B. Bischl, M. Lang, F. Pfisterer, N. G. Reich, J. Richter, P. Schratz, and R. Sonabend, *Mlr3 book*, Mar. 2021. [Online]. Available: https://mlr3book.mlr-org.com.

[161] N. A. Obuchowski and J. A. Bullen, "Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine," *Phys Med Biol*, vol. 63, no. 7, 07TR01, Mar. 2018. DOI: https://doi.org/10.1088/1361-6560/aab4b1. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29512515/.

[162] R. R. Bouckaert and E. Frank, "Evaluating the replicability of significance tests for comparing learning algorithms," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2004, pp. 3–12.

[163] D. C. Beidel and S. M. Turner, *Shy children, phobic adults : nature and treatment of social anxiety disorder*. Washington, D.C: American Psychological Association, 2007, ISBN: 9781591474524.

[164] J. A. Cooper, A. R. Arulpragasam, and M. T. Treadway, "Anhedonia in depression: biological mechanisms and computational models," *Curr Opin Behav Sci*, vol. 22, pp. 128–135, Aug. 2018. DOI: https://doi.org/10.1016/j.cobeha.2018.01.024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29503842/.

[165] T. D. Borkovec, O. M. Alcaine, and E. Behar, "Avoidance theory of worry and generalized anxiety disorder.," in *Generalized anxiety disorder: Advances in research and practice.* New York, NY, US: The Guilford Press, 2004, pp. 77–108, ISBN: 1-57230-972-5 (Hardcover).

[166] A. E. J. Mahoney, M. J. Hobbs, J. M. Newby, A. D. Williams, M. Sunderland, and G. Andrews, "The Worry Behaviors Inventory: Assessing the behavioral avoidance associated with generalized anxiety disorder," *J Affect Disord*, vol. 203, pp. 256–264, Oct. 2016. DOI: https://doi.org/10.1016/j.jad.2016.06.020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/27314812/.

[167] S. Place, D. Blanch-Hartigan, C. Rubin, C. Gorrostieta, C. Mead, J. Kane, B. P. Marx, J. Feast, T. Deckersbach, A. S. Pentland, A. Nierenberg, and A. Azarbayejani, "Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders," *Journal of Medical Internet Research*, vol. 19, no. 3, e75, 2017, ISSN: 1438-8871. DOI: 10.2196/jmir.6678.

[168] M. Boukhechba, P. Chow, K. Fua, B. A. Teachman, and L. E. Barnes, "Predicting social anxiety from global positioning system traces of college students: Feasibility study," *JMIR mental health*, vol. 5, no. 3, e10101, Jul. 4, 2018, ISSN: 2368-7959. DOI: 10.2196/10101.

[169] T. G. Nick and K. M. Campbell, "Logistic regression," *Methods Mol Biol*, vol. 404, pp. 273–301, 2007. DOI: https://doi.org/10.1007/978-1-59745-530-5_14. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18450055/.

[170] L. Iverach, R. G. Menzies, and R. E. Menzies, "Death anxiety and its role in psychopathology: Reviewing the status of a transdiagnostic construct," *Clinical Psychology Review*, vol. 34, no. 7, pp. 580–593, Nov. 2014, ISSN: 1873-7811. DOI: 10.1016/j.cpr.2014.09.002.

[171] D. Di Matteo, A. Fine, K. Fotinos, J. Rose, and M. Katzman, "Patient willingness to consent to mobile phone data collection for mental health apps: Structured questionnaire," *JMIR Mental Health*, vol. 5, no. 3, e56, 2018. DOI: 10.2196/mental.9539. [Online]. Available: https://mental.jmir.org/2018/3/e56/ (visited on 01/21/2020).

[172] T. P. Ryan, *Modern Experiment Design*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, 2007, ISBN: 9780471210771. DOI: http://dx.doi.org/10.1002/0470074353.

[173] R. Stewart and S. Velupillai, "Applied natural language processing in mental health big data," *Neuropsychopharmacology*, vol. 46, no. 1, pp. 252–253, 2021. DOI: 10.1038/s41386-020-00842-1. [Online]. Available: https://doi.org/10.1038/s41386-020-00842-1.

[174] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018. DOI: 10.1109/TASLP.2018.2842159.

[175] G. Sagl, B. Resch, A. Petutschnig, K. Kyriakou, M. Liedlgruber, and F. H. Wilhelm, "Wearables and the quantified self: Systematic benchmarking of physiological sensors," *Sensors*, vol. 19, no. 20, 2019, ISSN: 1424-8220. DOI: 10.3390/s19204448. [Online]. Available: https://www.mdpi.com/1424-8220/19/20/4448.

[176] M. Elgendi and C. Menon, "Assessing Anxiety Disorders Using Wearable Devices: Challenges and Future Directions," *Brain Sci*, vol. 9, no. 3, Mar. 2019. DOI: https://doi.org/10.3390/brainsci9030050.

[177] H. Lee, H. Ahn, S. Choi, and W. Choi, "The SAMS: Smartphone addiction management system and verification," *Journal of Medical Systems*, vol. 38, no. 1, p. 1, 2014. DOI: 10.1007/s10916-013-0001-1. [Online]. Available: https://doi.org/10.1007/s10916-013-0001-1.

[178] A. Grunerbl, A. Muaremi, V. Osmani, G. Bahle, S. Ohler, G. Troster, O. Mayora, C. Haring, and P. Lukowicz, "Smartphone-based recognition of states and state changes in bipolar disorder patients," *IEEE journal of biomedical and health informatics*, vol. 19, no. 1, pp. 140–148, Jan. 2015, ISSN: 2168-2208. DOI: 10.1109/JBHI.2014.2343154.

[179] D. Ben-Zeev, R. Wang, S. Abdullah, R. Brian, E. A. Scherer, L. A. Mistler, M. Hauser, J. M. Kane, T. Choudhury, and A. Campbell, "Mobile behavioral sensing in outpatients and inpatients with schizophrenia," *Psychiatric services (Washington, D.C.)*, vol. 67, no. 5, pp. 558–561, May 1, 2016, ISSN: 1075-2730. DOI: 10.1176/appi.ps.201500130. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4918904/ (visited on 01/03/2020).

[180] T. Dalgleish, M. Black, D. Johnston, and A. Bevan, "Transdiagnostic approaches to mental health problems: Current status and future directions," *J Consult Clin Psychol*, vol. 88, no. 3, pp. 179–195, Mar. 2020. DOI: http://dx.doi.org/10.1037/ccp0000482.

[181] J. Kaufman and D. Charney, "Comorbidity of mood and anxiety disorders," *Depression and Anxiety*, vol. 12, no. S1, pp. 69–76, 2000. DOI: https://doi.org/10.1002/1520-6394(2000)12:1+<69::AID-DA9>3.0.CO;2-K.

[182] N. Palmius, K. E. A. Saunders, O. Carr, J. R. Geddes, G. M. Goodwin, and M. D. Vos, "Group-personalized regression models for predicting mental health scores from objective mobile phone data streams: Observational study," *Journal of Medical Internet Research*, vol. 20, no. 10, e10194, 2018. DOI: 10.2196/10194. [Online]. Available: https://www.jmir.org/2018/10/e10194/ (visited on 01/09/2020).

[183] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation*, M. Agrawal, D. Du, Z. Duan, and A. Li, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 1–19. DOI: https://doi.org/10.1007/978-3-540-79228-4_1.