March 2018

Sensing Depression

Ada Dogrucu
Worcester Polytechnic Institute

Aleksa Perucic
Worcester Polytechnic Institute

Anabella Isaro
Worcester Polytechnic Institute

Damon Clark Ball
Worcester Polytechnic Institute

Sensing Depression

A Major Qualifying Project Proposal

Submitted to the Faculty of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Bachelor of Science

In

Computer Science

By

Written By:

Damon Ball

Ada Dogrucu

Anabella Isaro

Alex Perucic

Advised By:

Professor Elke Rundensteiner

Professor Emmanuel Agu

Date: Mar 23, 2018

Acknowledgements

This Major Qualifying Project could not have been completed without the advice and guidance of Professor Elke Rundensteiner and Professor Emmanuel Agu of the Computer Science Department at Worcester Polytechnic Institute. We would also like to thank Professor Craig Wills, Computer Science Department Head at Worcester Polytechnic Institute, for authorizing funds for our project, without which our research wouldn't have been possible. We also would like to thank Ermal Toto, whose advice on machine learning proved to be critically helpful throughout the duration of our project.

Abstract

According to the Center for Disease Control (CDC), more than 1 out of 20 Americans of age 12 or older have suffered from depression between the years of 2009 and 2012. The hallmark indicator of depressive disorders is a presence of “sad, empty, or irritable mood, accompanied by somatic and cognitive changes that significantly affect the individual’s capacity to function”. Diagnosing depression can be difficult and unreliable, as current tools of diagnosis such as the PHQ-9 questionnaire require the patient to have a strong capacity of introspection in order for the results to be accurate. The overall goal of our project is to provide a tool for doctors to effortlessly detect depression, and in effect, achieve greater coverage in detecting depression over the general population.

We use machine learning techniques to create a mobile application that infers a smartphone user’s severity of depression (or lack thereof) from data scraped off their phone and social media websites, which includes GPS data, call and text metadata, social media usage data and voice samples. The information is collected from the prior two weeks from whenever a subject initiated an assessment, and the assessment is done on the spot, providing instant feedback to the doctor and the patient.

This work is novel because unlike prior approaches, user data is obtained on the spot, as opposed to being actively accumulated over a period of time. Through our study, we have demonstrated the feasibility of this approach to diagnosing depression, achieving an average test set RMSE of 5.67 across all modalities in the task of PHQ-9 score predictions.

The Goal of this MQP

The use case for our project is that our app can be used for depression screening instead of paper questionnaires, increasing the accuracy and convenience of screening. Any patient simply can download our application and then will receive a diagnosis on the spot. This presents us with the unique challenge of being able to exclusively use data that a smartphone (or the user's social media accounts) have already gathered in prior 2 weeks or can instantaneously gather at the time of the assessment.

The types of data that fit within this definition are further divided into active and passive modalities. Active modalities require input from the user at the time of the assessment, such as recording a voice sample. Passive modalities include GPS and SMS data, which can be passively scraped in the background without user input.

We aim to maximize the usage of passive modalities, while minimizing the usage of active modalities in order to design the lowest burden application for the user that fits our use case.

Table of Contents

List of Figures

List of Tables

1. Introduction

According to the Center for Disease Control (CDC), more than 1 out of 20 Americans of age 12 or older have suffered from depression between the years of 2009 and 2012. A common feature of depressive disorders is a presence of "sad, empty, or irritable mood, accompanied by somatic and cognitive changes that significantly affect the individual's capacity to function."(DSM-V, 2013). Depressive disorders, as they are medically labeled, come in various types. Common examples are disruptive mood regulation disorder, major depressive disorder and persistent depressive disorder (DSM-V, 2013). The symptoms for depression commonly include loss of enjoyment in things one previously enjoyed, social withdrawal, exacerbation of pre-existing pains, fatigue, agitation, and diminished activity.

According to the British Psychological Society, people with depression are four times more likely to commit suicide, which accounts for two-thirds of all suicides (British Psychological Society 2010). In the United States, suicide is a major public health concern and is the third leading cause of death among young adults aged 18–24 years (Scottye et. al. 2009). Depression can also lead to other physical diseases such as heart disease, type 2 diabetes and obesity and can subsequently accelerate the onset of mental decline, make a person more likely to abuse substances, and even cause changes in their immune system which makes them more susceptible to cancer (Goodwin et. al, 2006). In addition to the deleterious health-related effects of depression, there is also a financial cost related to depression. The cost of handling depression in the US was $210.5 billion in 2010 (Greenberg et. al. 2010). 50% of this cost was attributed to workplace costs, particularly by presenteeism (reduced productivity while at work) and absenteeism (missed days from work). The rest was divided amongst direct medical costs, which account for 45% of the cost, and suicides, which account for 5% of the cost. Emotional and motivational effects of depression also decrease a person's ability to work effectively and can cause losses in one's income (British Psychological Society, 2010).

Despite the prevalence and severity of depression, it is one of the major mental health disorders that is least diagnosed. Diagnosis and assessment of depression symptoms relies almost entirely on information provided by patients, family members, peers or caregivers (Yang et al., 2013). However, this form of report is unreliable since it depends on complete honesty from the reporter. Self and perceived stigma surrounding depression is prevalent in communities all over

the world, and is associated with a reluctance to seek professional help (Kathleen M et al, 2014). In a study conducted in the UK, of the 130 subjects that were known to exhibit symptoms of depression, only 80 had consulted their general practitioner. The reasons behind their inability to confer with their GP on this issue included feeling too embarrassed to discuss their depressive symptoms with anyone, and thinking nobody could help (British Psychological Society, 2010). The stigma around mental illness prevents people from getting treatment in order to avoid the public label of mental illness and not feel shame and guilt about themselves (Brown et. al. 2010). Patients are often reluctant to share their depressive feelings with doctors and therefore a discussion about depression often depends largely on a general practitioner's ability to communicate with the patient (British Psychological Society 2010).

A lot of research has gone into alternative methods of detecting depression. This body of research includes studies that try to correlate depression with various types of data including voice patterns, facial expressions, locomotion, sleep patterns, and text post data. Studies have shown that tracking people's movement patterns can accurately be used to detect depression (Canzian et. al. 2015 and Saeb et. al. 2016). Other studies show that analyzing social media users' posts and interaction habits can detect depression. The studies collected the public posts of these users and used machine learning to classify the messages into a depression rating (Park et. al. 2012 and Park et. al. 2013). Another seminal paper uses algorithms on both vocal prosody and facial expressions to detect depression. (Cohn et. al., 2009)

Today, 77% of all Americans own a smartphone (Smith 2017) that is equipped with a variety of sensors. These sensors make it possible for a smartphone to track many different types of data including GPS location, call habits, text habits, and sleep patterns. Our goal is to create a mobile application that can detect depression by grabbing these types of data from a patient's smartphone. This depression meter could then potentially replace the more subjective paper based methods of screening. The application is designed to give a patient entering the ER an on-the-spot depression rating (how severe depression symptoms they show). This presents the unique challenge of creating a high-quality depression screen exclusively based on data that a smartphone has already gathered and that can be accessed by our application at the time the patient arrives at the ER. The overall goal of our project is to provide a tool for doctors to more easily detect depression and better yet achieve greater coverage in detecting depression over the general population.

2. Literature Review

2.1. Background on Depression

The World Health Organization (WHO) ranked depression amongst the most disabling illnesses affecting the world's population. According to the WHO, more than 300 million people around the globe suffer from depression. In the United States alone, more than 16 million Americans aged 18 years or older suffer from depression (NIH, 2015). In context, depression refers to a wide range of mental health challenges that are associated with emotional, cognitive, physical, and behavioral symptoms, such as loss of enjoyment and interest in ordinary things, the lack of positive affect, or distinguished mood changes (DSM, 2013). These depressive symptoms can be disabling and pervasive, impacting not only the individual but also their families and the community in which the individual resides at large. The government has invested in research related to depression treatment in an effort to reduce the effects of depression, however despite the increasing availability and variety of treatment, under diagnosis and under treatment of depression are still a major problem. In the following paragraphs, we will review literature on certain demographics that are highly affected by depression, and the various screening methods for depression.

2.1.1. Influence of Age and Gender on Depressive Conditions

The prevalence of depression cuts across race, gender, age and socioeconomic status. According to National Alliance of Mental Illness, young adults between ages of 18 - 25 are 60% more likely to suffer from depression than people aged 50 or older. Clinical cohort studies have confirmed that the rates of depression rise sharply after puberty with immediate and long term risks. However the rates of under diagnosis and under treatment are particularly high in young adults compared to any other age group (Thapar et al. 2010). Without a diagnosis or treatment, mild depression can develop into a more severe form of depression, adversely affecting the schooling, educational attainment, and relationships of the affected individual (British Medical Journal, 2010). Depression among elderly is also a serious public health concern that is yet to be addressed (Chapman and Perry, 2010). This concern, of course, is in relation with the decrease in social and physical abilities among the elderly due to the decline in their cognitive abilities. Research has shown that there is a strong relationship between physical activity and depression,

where the more active an individual is the less likely they are to become depressed (Chodsko-Zajko et al. 2011, Nelson el at., 2007).

Furthermore, women are consistently reported to have a larger incidence of depression disorder than men. In 2010, the global annual prevalence of depression between women and men was 5.5% and 3.2%, respectively, representing a 1.7-fold greater incidence in women (Albert, 2015). It is important to note that this finding on the women to men prevalence ratio globally suggested that the differential risk stems mainly from the biological sex differences and depend less on other numerous confounding economic or social factors. In a National Health Population Survey, the data showed that the rates of depression peak amongst women during the reproductive age (ages 15 to 44) (Stewart et al., 2004). It is reported that during pre-pubescence the rates of depression between both genders are equal (Bebbington et al. 2003). However, the gender gap rates emerge at puberty and decline after menopause. Where at ages of above 65 years, both men and women show a decline in depression rates, and the prevalence becomes similar for both genders. This highlights the complexity that contributes to the psychological and biological factors of depression.

2.1.2. Forms of Depressive Disorders

According to the DSM-5, forms of depressive disorders include disruptive mood dysregulation disorder, major depressive disorder, persistent depressive disorder, premenstrual dysphoric disorder, and substance/medication-induced depressive disorder (DSM-V; American Psychiatric Association, 2016). Although these different disorders differ in duration, timing and presumed etiology, the common features commonly are a presence of sadness, loss of enjoyment and irritable mood, accompanied by cognitive changes. After reviewing different types of depression, our project decided to focus on predicting Major Depressive disorder, since according to our finding it is the most prevalent one across all spectrum in terms of age and gender. Major depression in comparison with other forms of depression as mention also extends for a longer period up to 2 years.

2.1.2.1. Disruptive Mood Dysregulation

Disruptive Mood Dysregulation is the most common type of depressive disorder amongst children aged 12, and it is characterized by persistent irritability and frequent episodes of extreme behavioral dyscontrol (DSM-5; American Psychiatric Association, 2016). These

extreme behaviors can manifest into verbal rage such as severe recurrent temper outburst. They also do not often correlate with developmental level and occur three or more times a week on average. The overall prevalence of disruptive mood dysregulation disorder is about 2 - 5% over the periods of 6-months and 1-year, with the majority of children presented to clinics with features of disruptive mood dysregulation being predominantly male (DSM-5; American Psychiatric Association, 2016). It is important to note that disruptive mood dysregulation disorder needs to be carefully distinguished from other related mental health conditions.

2.1.2.2. Major Depressive Disorders

Major depressive disorder is often associated with considerable morbidity, disability and an elevated risk of suicide. It is often categorized into two types of chronic and non-chronic major depressive disorders; individuals with chronic and non-chronic depression differ on a large spectrum of clinically and etiologically variables such as comorbidity, impairment, and psychopathology. Chronic depression refers to a more severe condition that presents itself with the symptom of recurrent depressive episodes that occur continuously over a period of 2 years, while non-chronic depression is often associated with less severe impairments and risk factors (Klein et al., 2006). Individuals with chronic depression have a higher rate of comorbid conditions and more extreme personality traits such as neuroticism and higher levels of suicidality rates (Rios et al., 2003). According to the Diagnostic and Statistical Manual of Mental Health Disorders, the prevalence of major depressive disorder that persists for twelve months in the United States is approximately 7% with specific differences in age groups. It may appear at any age, however, the likelihood tends to peak at puberty. Many dysfunctional consequences of major depressive disorder develop from individual symptoms.

2.1.2.3. Persistent Depressive Disorder

In the United States, the 12-month prevalence of persistent depressive disorder is approximately 0.5% (DSM-5; American Psychiatric Association, 2016). The concept of temporal prevalence refers to the proportion of a population that has the condition at some time during the specified amount of time, in this case, 12 months. persistent depressive disorder is a consolidation of chronic major depressive disorder and dysthymic disorder, however they differ in recurrence. Note that persistent depressive disorder can occur during a major depression and

vice versa (DSM-IV; American Psychiatric Association, 1994). However, the criteria of current episodes of major depressive disorder superimposed on existing persistent depressive disorder (Nemeroff, Charles B., et. at., 2003). According to the DSM-5, the essential feature of persistent depressive disorder is that it occurs for most of the days for at least two years for adults and one year for adolescents. Some of the symptoms of persistent depressive disorder are similar to general major depression symptoms such as low energy, feelings of hopelessness, insomnia or hypersomnia. However, some symptoms are unique to persistent depressive disorder and can cause clinically significant distress or impairment. While symptoms of persistent depressive disorder are often on par with that of major depressive disorder, these symptoms are likely to subside.

2.1.2.4. Premenstrual Dysphoric Disorder

Up to 8% of women in their reproductive years suffer from premenstrual dysphoric disorder, a more severe form of premenstrual syndrome that affects about 50 - 80% of women in their menstruation cycle (Khazaie, et al. 2016). Premenstrual syndrome differs from premenstrual dysphoric disorder and only 2 to 8% of women meet the strict criteria of cognitive-affective, behavioral and physical symptoms as defined in both DSM-4 and DSM-5. Essential features of premenstrual dysphoric disorder comprise of cognitive symptoms such as irritability, dysphoria or affective lability and physical symptoms such as muscle pain or bloating and behavioral symptoms such as poor concentration, decreased interests, lethargy, and changes in sleep during the premenstrual cycle (Andrade, 2016). However, according to the DSM-5 onset premenstrual dysphoric disorder symptoms can also occur at any point after menarche. Symptoms of premenstrual dysphoric disorder often result in clinically meaningful distress and unlike premenstrual syndrome, it cannot be induced by the use of steroidal contraceptives.

2.1.2.5. Substance/Medication-Induced Depressive Disorder

In United States, 0.26% of the national representative of the adult population suffers from substance/medication-induced depressive disorder during the period of a lifetime (DSM-5; American Psychiatric Association, 2016). Substance/medication-induced depressive disorder refers to the specificity of the substance causing the depressive symptoms. As defined by DSM-5, the diagnostic features of substance/medication-Induced Depressive disorder symptoms

include those of any depressive disorder, such as major depressive disorder; however, the symptoms are associated with the inhalation, ingestion or injection of a substance (e.g. drug of abuse, toxin, psychotropic medication, other medication). Some medications that affect the central nervous system or are used as stimulants are likely to induce depressive mood disturbance, however, clinical judgment is essential to determine whether the medication is truly associated with inducing the depressive disorder (DSM-5; American Psychiatric Association, 2016).

2.1.2.6. Depression Screening and Diagnosis

Despite the high prevalence of depressive disorders, studies have shown that the detection of depression in primary care setting is still suboptimal. In an article published by the British Medical Journal, it was observed that clinicians often overlooked signs of depression; clinicians failed to recognize depression in approximately 30-50% of the cases (Josefson, 2002). The US Preventive Services Task Force now recommends depression screening for clinical practices to assure accurate diagnosis with greater coverage. The task force based its recommendation on literature from the Medline databases; Medline study included fourteen randomized trials that examined the effect of screening in enabling accurate identification and treatment of diseases. Seven of the studies showed that incorporating screening for depression into normal care routine increased the diagnosis of depression by a factor of 2 to 3.

It is important to understand the various tools used for depression screening. These tools include Hamilton Depression Rating Scale (HDRS), Beck Depression Inventory (BDI), Patient Health Questionnaire (PHQ), Major Depression Inventory (MDI), Center of Epidemiologic Studies Depression Scale (CES-D), Zung Self-Rating Depression Scale (GDS), Geriatric Depression Scale (GDS), Cornell Scale of Depression in Dementia (CSDD). Depression screening is a standard part on checking in at the hospital and most specifically in the Mental Health Service ER. In Table 1, a comparison is drawn between, time, cost, specificity and sensitivity of the major screening methods reported, this will help infer the screening method that is better suited for project.

2.1.2.7. Clinically Validated Measures of Depression

Amongst a myriad of clinically validated measures of depression, we surveyed each method and decided to use PHQ-9 for its relative simplicity, accuracy and prevalence (Pfizer,

1999). In this section, we present our research on the prominent clinically validated measures of depression, and include a quick overview of how the particular measure would potentially fit within the structure of our project.

For all these measures we include a score of specificity and sensitivity. In a clinical context, these concepts refer to the concept of true negative rate and true positive rate respectively. Specificity measures the proportion of negatives that are correctly identified. Sensitivity measures the proportion of positives that are correctly identified.

2.1.2.7.1. Hamilton Depression Rating Scale (HDRS)

Hamilton Depression Rating Scale is the most widely used interview scale. HDRS specializes in measuring the severity of depression and has also proved useful in determining the patient's level of depression before, during and after treatment. It comprises structured interview guides, self-report forms and computerized versions of the HDRS form (Bienenfeld et al, 2016). In general, it is considered to be the most suitable for inpatient patients. The Hamilton Depression Rating Scale form consists of large items concerning a number of somatic symptoms and a few cognitive symptoms and generally take 15-20 minutes to complete. The HDRS form lists 21 items and the scoring base is on the first 17 items with 18 to 21 items used to further qualify depression. The scoring is on a scale of 5 for eight items, ranging from 0= not present to 4 = severe and the nine other items are scored on scale of 2. The sum of the scores are used to determine between normal to very severe depression; with a total score between 0 - 7 = normal, 8 - 13 = mild depression, 14 - 18 = moderate depression, 19 - 22 = severe depression, and >= 23 = very severe depression.

The HDRS presents a comprehensive measure of depression, yet for data collection, it’s length presents a challenge concerning the length of our data collection survey. Through our research we have concluded that engaged participants in a study provide more correct data, thus we will not use the HDRS.

2.1.2.7.2. Beck Depression Inventory (BDI)

The Beck Depression Inventory was developed by Aaron Beck and it is based on symptoms he observed as common amongst patients suffering from depression. The BDI is most commonly used on a self-report rating scale that measures characteristic attitudes and symptoms

of depression (Bienenfeld et al, 2016). It consists of behavioral, emotional and somatic symptoms and take between 5-10 minutes to complete. The Beck Depression Inventory form also contains 21 items; the items are scored on a scale of 3, ranging from 0 = not present and 3 = severe. The sum of the score are to be used to determine between normal and severe; with a total score between 0 - 10 = normal, 10 - 18 = mild depression, 19 - 29 = moderate depression, and >= 30 = severe depression. There exists various versions of BDI including BDI-II, which employs the same scoring scale as BDI with different cutoffs, and BDI-PC which is a 7 item scale for primary care outpatients.

The BDI presents a short, robust but not popular measure of depression. We forego this method for it's relative lack of infamy.

2.1.2.7.3. Patient Health Questionnaire (PHQ)

The patient Health Questionnaire is a multipurpose instrument for screening, monitoring, diagnosing, and measuring severity of depression (Bienenfeld et al, 2016). It is a self-administered tool that has two different forms, the PHQ-2 that contains 2 items and the PHQ-9 that contains 9 items. PHQ-2 assess the frequency of depressive episodes and anhedonia over the past two weeks, while the PHQ-9 establishes a clinical diagnosis of depression and tracks the severity of the symptoms. PHQ- is scored on a scale of 3, with 0 = not present, and 3 = severe. A PHQ-2 score higher than 3 gives a sensitivity of 83% and a specificity of 92% of major depression; while the cut point for PHQ-9 is greater or equal to 10 gives a sensitivity of 88% and a specificity of 88% of major depression. The sum of the scores is then used to determine between normal and severe depression. A total score between 0 - 5 = mild depression, 6 - 10 = moderate depression, 11 - 15 = moderately severe depression, and 16 - 20 = severe depression. See Appendix A for the full set of the PHQ-9 questions.

The PHQ-9 presents a short, robust, popular and with its specificity and sensitivity, a proven way to infer depression through a questionnaire. We chose to use PHQ-9 in our project for these reasons.

PATIENT HEALTH QUESTIONNAIRE-9 (PHQ-9)

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
9. Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3

Figure 1: The PHQ-9 Questionnaire

2.1.2.7.4. Major Depression Inventory (MDI)

The major depression Inventory was developed based on the universal symptoms in DSM-IV major depression and ICD-10 moderate to severe.it is a self-rating scale that is used in the diagnosis and measurement of depression. As a diagnosis tool, there are 10 items with 1 = present and 0 = absence for each symptoms and as a measuring tool, the items are on a scale of 5 with 0 = not present and 5 = severe. The sum of the score are to be used to determine between normal and severe depression, the cutoff score is 26 at a maximum of 50. MDI algorithm has a sensitivity between 82% and 92%, with a specificity between 82% and 86%. For a diagnosis of major depression item 1 o 2 should be present among the 5 of 9 items presents, in this case items 4 and 5 are combined.

The MDI presents a more technically obfuscate, and a little less precise version of the PHQ-9. For this reason we will not utilize it in our project.

Screening Tool	Number of Items	Scoring Time	Psychometric Properties	Cost and Development
HDR	21 items	15 to 20 min	Sensitivity: 93% Specificity: 98%	Free
BDI	21 items	5 to 10 min	Sensitivity: 84% Specificity: 81%	Proprietary ($115/kit)
PHQ-9	9 plus severity items	<5min	Sensitivity: 88% Specificity: 88%	Free with permission
MDI	10 items	<5min	Sensitivity: 86% Specificity: 82%	Free

Table 1 - Mental Health Screening and Assessment Tools for Primary Care

As explained in the conclusion paragraphs in each section above, we opted to ascribe importance to scoring time more than anything, to minimize the amount of time a participant would spend completing the screening, and thusly to maximize the quality of data we receive. We then considered the psychometric quality of each questionnaire to decide on the PHQ-9 as our preferred method for measuring depression severity.

2.2. Detecting Depression Using Mobility Data

The recent popularity of smartphones allows any smartphone owner to view their GPS coordinates with unparalleled accuracy. Smartphones now possess the ability to track the places a person has gone, how much a person travels, and what type of transportation they were taking. Studies show that depression can be detected using Global Positioning System (GPS) data to track a patient's location and movements. Two studies, one by Canzian and one by Saeb, gathered data using software that subjects installed on their phones (Canzian et. al 2015, Saeb et. al. 2016). The software recorded subjects' location over a period of time. The study conducted by Saeb sampled data every 5 minutes for 10 weeks, while the study by Canzian recorded locations at variable rates of 5 minutes or more for 2 weeks . However, since our application must use existing data found on the phone, sample rate and period will entirely depend on the

available sources of GPS data stored on the phone and will look backwards at already stored data instead of looking forward.

Both studies converted the raw GPS data into a set of similar metrics. The metrics studied between the two studies included total distance traveled, maximum distance between two locations, "radius of gyration" or the average distance from the center point, standard deviation of displacements, number of places visited, number of different significant places visited, time spent at home, and difference from routine. The raw data is grouped into "stop places" which are places the person stopped at for a certain threshold of time, as described in the study by Canzian. The stop places consist of an ID, the time spent at the location (ti = time of arrival, td = time of departure, and the latitude and longitude of the location (C) (Canzian et. al 2015).

$$Pl = (ID, ti, td, C)$$

Total Distance: Total distance is calculated by adding up the differences between the latitude and longitude of all successive stop places. This is a basic measurement taken by both studies, and can be used as a general measure of how much a person is moving around (Saeb et. al. 2016).

$$Total\ Distance = \sum\sqrt{((lati - lati - 1)^2 + (longi - longi - 1)^2)}$$

Maximum distance between locations: Maximum distance is simply the largest change in position between two stop locations. A smaller maximum distance could indicate a person is less willing to move as far (Canzian et. al 2015).

Radius of Gyration: In the study by Canzian, radius of gyration is the term used for the average distance from the center of all locations for a given period. Radius of gyration is found by multiplying the square of the distance from each location to the center by the time spent at that location, then adding the result for each location together. The square root of this number divided by the total time spent at the locations it the radius of gyration. This is another metric for measuring how far the subject is traveling from their home or other places they spend most of their time (Canzian et. al 2015).

$$G(t1,t2) = \sqrt{((1/T)\sum(t \cdot d(Ci,Ccen)^2))}$$

Standard Deviation of displacements: Standard deviation in the distances traveled is another metric used to determine how much the person is moving around, similarly to radius of gyration but does not take into account where the center is for a subject (Canzian et. al 2015).

$$\sigma dis = \sqrt{(\,(1/N-1)\sum(\,d(Ci,Ci+1) - Average\ Displacement)^2\,))}$$

Number of places visited: Number of places visited is the number of "stop places" the raw data was grouped into, which as stated before were places the subjects stopped at for a certain amount of time. A high number would indicate a very busy, energetic person (Canzian et. al 2015).

Number of different significant places visited: Significant places are the top visited places. In the study by Canzian, the top 10 most visited places were chosen to be significant, but for the purposes of the app a different more suitable number may be chosen instead to better suit the data. This metric is useful for seeing where a person is spending their time, whether public or private or in a spot that is likely to cause depression (Canzian et. al 2015).

Time spent at home: The time spent at home would simply be the percentage of time spent at the stop place corresponding to the subject's home. A lot of time spent at home could possibly represent a lethargic subject (Saeb et. al. 2016).

Difference from routine: The difference in routine is a measure of how many stop places are not part of the subject's daily routine. The difference in routine would be determined by comparing the "stop places" a subject visits each day and what times they visit the places, and measuring how much this changes day to day. A high difference from the routine for a day would signify the person was very active and went a lot of places that they do not typically go (Canzian et. al 2015).

Each metric is a different way to measure the activity level of a person. If a person is less active than usual, it could be a sign of a loss of enjoyment in normal activities, social withdrawal, or possibly desiring less movement per exacerbated pains, which are known symptoms of depression (The British Psychological Society 2010). It is also important to note that in the study by Saeb, results taken from weekend days were found to be stronger indicators for depression than weekdays likely due to the increased choice a person has outside of normal work days (Saeb et. al 2016).

Each metric was measured for each subject over the course of the study and compared to the subject's results in the PHQ test. In the Canzian study, the PHQ test was self-administered each day, and in the Saeb study the test was administered at the beginning and at the end of the trial. The trends in each subject's GPS metrics were compared with their change in PHQ scores over time to look for correlations. Both studies found a strong correlation between the metrics and PHQ depression scores. Both also found that the higher level metrics, like location variance, number of different places, and home stay had a particularly high correlation. However, the study by Canzian found that the low level metrics, such as distance traveled and longest distance between locations, were the strongest when the data was gathered over a longer period of time. However, the application we designed would gather data from an existing database which limited us to only looking back two weeks. Therefore, our results using Canzian and Saeb's methods may not be as strong as ones that gathered data further back into the past.

2.3. Detecting Depression Using Text Messages and Social Media

According to the Pew Research Center, 73% of all Americans send and receive text messages (Aaron Smith, 2011). Social media sites where users can post textual messages are also extremely popular, with 79% of Americans using Facebook, and 24% of Americans using Twitter (Greenwood et. al. 2016). Studies show that analyzing a person's text posts for the sentiments of the posts, the frequency of posting, and linguistic styles of the posts can reveal their mood. A study by Ramon Rodrigues showed that text posts on Facebook can be analyzed to detect the mood of the user (Rodrigues et. al. 2016). In the study Rodrigues gathered a user's' posts and analyzed each individual post using a sentiment analysis tool. A sentiment analysis is a tool that analyzes a piece of text, either at the document level or the sentence level, and forms an "opinion" on if the text conveys a positive, negative, or neutral sentiment. The tool can also

attempt to label the object of the sentiment, allowing it to detect multiple, possibly opposite sentiments in one sentence. The study tested multiple different sentiment analysis tools to determine which were most accurate in detecting the emotional state in cancer patients. The result found that adapting the methods that these sentiment analysis tools used to create a custom tool for the desired job was more effective than using existing tools. The author believes that this method could be very effective for detecting other diseases such as depression (Rodrigues et. al. 2016). Therefore, our application could use machine learning to detect patterns useful for detecting depression that could then be used to adapt an already existing sentiment analysis tool to detect for depression.

Another study conducted by Minsu Park analyzed posts from Twitter to detect depressive moods (Park et. al. 2012). Park gathered a group of 69 participants and acquired permission to collect their tweets. Each subject was given the self-administered Center for Epidemiologic Studies Depression Scale (CES-D) test to determine their position on the depression scale. From the results of the test, 41 subjects scored low or mild depression and was used as the "normal" group, and 28 participants scored positive for depression and were used as the "depressed" group. The tweets of all participants were analyzed using the Linguistic Inquiry and Word Count (LIWC) sentiment tool to categorize tweets into psychologically meaningful categories. The LIWC contains a dictionary of several thousand words with each word having a score in six different criteria: social, affective, cognitive, perceptual, biological processes, and relativity (Park et. al. 2012). Each criterion has several categories and subcategories, and if a word is found in the tweet belonging to a category, that word's score will be incremented. The conductors of the study plotted the CES-D scores against each category of the LIWC tool to see which categories correlated with a high depression score. They found that depressed users tended to tweet about themselves more than interact with other users, and that these tweets often contained words in the "affect" category of the LIWC tool in subcategories such as "anger", "causation", and "friends." Our application could apply the LIWC tool on the data it gathers to look for high scores in these categories. The study also found that Twitter users are very willing to post private information about themselves on Twitter, such as history with depression or treatment, meaning for the purposes of this application it would not be difficult to find users that have a stated history of depression and use their posts as training data to look for patterns.

A third study set out to predict postpartum depression in new mothers before they show any sign by analyzing their Twitter use (Choudhury 2013). This study looked at four measures: engagement, ego-centric, emotion, and linguistic style.

Engagement: Engagement is measured by the volume of tweets, replies, and retweets per day. The average number of posts over time reveals whether a mother is more or less active on social media than before. The number of replies is a measure of more direct social interaction the person is engaging in. Finally, retweets indicates a person's contribution to sharing information in their social network.

Ego-centric: How ego-centric a user is measured by how many followers a person has and how many people a person follows. The number of followers a person gains or loses per day shows how popular the person is on social media, while how many people a person follows measures how active they are socially or how much they are actively attempting to stay connected.

Emotion: For measures of emotion this study uses the same sentiment analysis tool as the study by Park, the LIWC tool. In fact, the postpartum study also focused on the same negative affect categories of negative emotions including anger, anxiety, and sadness.

Linguistic Style: Linguistic style measures which types of words the person chooses to express themselves with, whether they be articles, auxiliary verbs, personal pronouns, prepositions etc. The style chosen can show the behavioral characteristics due to their social environment.

All four categories were found to be effective at predicting postpartum depression, each with approximately 70% accuracy. Using the four categories of measurements, the study used different designs to form predictions as to which mothers would develop postpartum depression. The results of each design's predictions were compared with the actual results of which mothers developed postpartum depression to see which was most effective.

2.4. Detecting Depression through Instagram

In their study, Reece et. al. have demonstrated that it is possible to detect depression using Instagram information of a user with a success rate that outperforms the medical general practitioners' average diagnostic success rate for depression (Reece et. al, 2017). In the context of this study, Instagram information refers to features extracted from a user's Instagram account. The information gathered includes how many comments and likes a post has, the Hue Saturation Value (HSV) of individual photos, posting frequency of the user and the number of faces in every picture. Their model found only about a third of the actual depressed observation with 31.8% recall, 83.3% specificity and 54.1% precision.

An interesting fact to note is that, the rate of success achieved by the study still holds up when the analysis is restricted to a user's Instagram posts made before that user was first diagnosed with depression. In other words, Reece et. al. have demonstrated that it is possible to detect depression in the retroactive way our paper aims to do.

In the study, featurization was conducted using machine vision and post metadata. The classification was done using a 1200-tree Random Forests classifier. Researchers also provide statistical analysis for individual predictors using logistic regression to determine the strength of every single predictor.

2.4.1. Face Count

For machine vision, researchers used a face recognition algorithm based on OpenCV to determine the number of faces in each post. Their findings presented intuitive results; having more posts of single faces posted correlated with depression, whereas in having more faces in a post negatively correlated with depression. In simple terms, depressed individuals tend to have more photos of singular faces posted, and not a lot of posts with many faces posted. The researchers deem it likely that having more faces in posts portrays a strong social life with many social interactions, since the number of social interactions one has is negatively correlated with depression, and that a large number of faces in Instagram posts are indicative of a strong social life and support system.

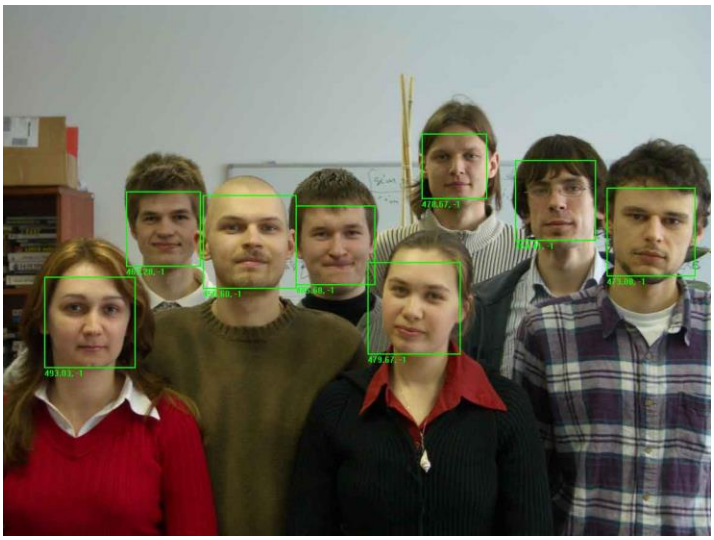

Figure 2: Face count as a possible metric for social life quality

2.4.2. Picture HSV and Filters:

For filters, the study tried to positively or negatively correlate the use of every single filter Instagram offers with depression. While many of the filters showed no statistically significant correlation, the “Inkwell” and “Valencia” filters stood out, positively and negatively correlating with depression respectively. Notice in Figure 3 that Inkwell reduces every pixel to grayscale whereas in Valencia gives everything a redder, softer hue.

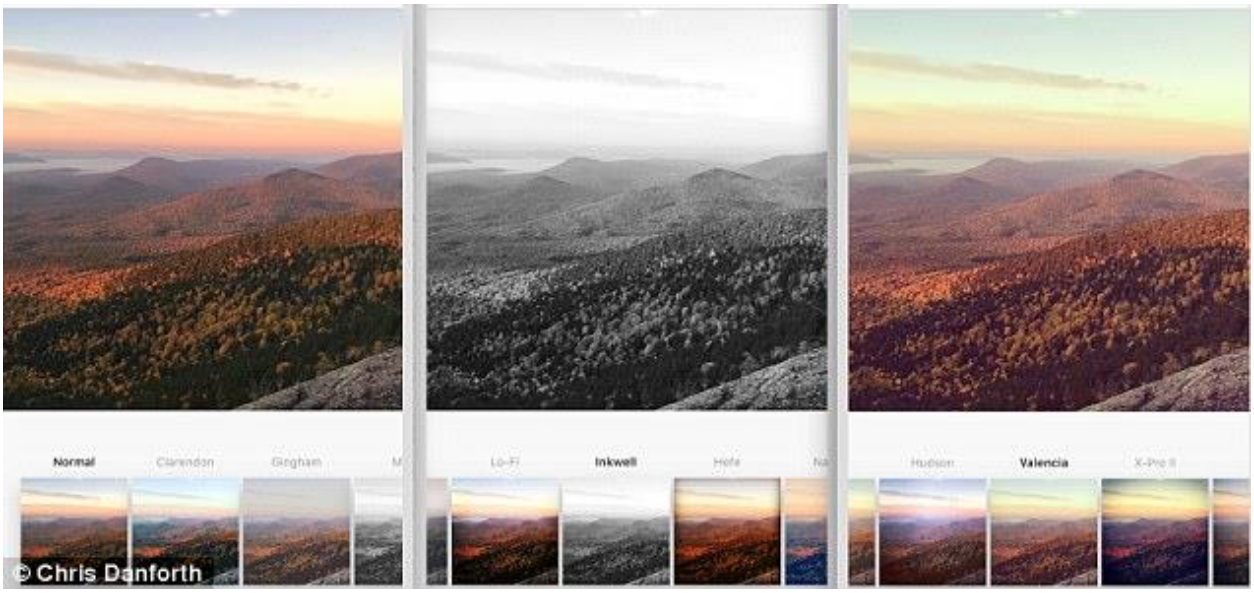

Figure 3: Filters Normal, Inkwell, and Valencia.

For image analysis, the researchers brought under scrutiny the hue, saturation and brightness levels of Instagram posts. Through averaging pixel values, they arrived at a hue,

brightness and saturation levels for every picture. They found increased hue, decreased saturation and decreased brightness to be predictors of depression.

2.4.3. User Interaction:

Instagram post metadata proves to be a good predictor of depression as well. In the dataset, more comments and lower number of likes positively correlated with depression. (Reece et. al, 2017) note that one way in which this study can be improved upon is by factoring in the contents of comments under a post.

2.5. Detecting Depression from Sleep Patterns

Sleep duration has been shown to be a very accurate predictor of depression. This coincides with traditional research on depression, where one of the most common symptoms of potential depression is lack of sleep (Nutt, 2008). Anxiety tends to accompany depression, and an anxious state is known to put an individual in a heightened emotional state, which makes it difficult to calm down enough to sleep. It is hypothesized that conditions such as insomnia are a symptom of untreated depression and/or anxiety.

A Dartmouth StudentLife study looked at the sleep duration of students and how this related to their emotional state. It examined smartphone-sensed data from 48 students over the course of 10 weeks and found a significant correlation between depression and sleep duration ($r = -0.382$, $p = 0.020$). The study used a sleep classifier based on their previous work. The phone would infer sleep data using light features, phone lock state, activity features, and sound features from the microphone. Their sleep model would combine all these different features in order to get a good estimate of sleep duration. It had a 95% prediction accuracy, with a +/- 25 minute margin of error.

An interesting conclusion was that as the term progressed, the students got less and less sleep, which lead to higher and higher rates of depression found on the PHQ-9. The researchers found it difficult to determine the cause and effect of this relationship, since it could have been that the students were sleeping less due to having schoolwork, and then this would have caused depression, or it could have been that the student's' already existing depression was amplified by the increase in stress.

2.6. Sleep Detection

As previously mentioned, the StudentLife researchers used four sensors to determine duration of sleep: light features, phone usage features including phone lock state, activity features (e.g. stationary), and sound features from the microphone. The individual features on their own were weak classifiers. However, when combined, these four features proved to be very accurate in predicting sleep duration, with a 95% prediction accuracy (+/- 25 mins of the ground truth).

2.6.1. Data and Results

The StudentLife researchers collected a total of 52.6 GB of data over the course of 10 weeks. They administered the PHQ-9 questionnaire at the start and the end of the 10 week period (pre and post). Apart from sleep detection, they also examined: activity data (indoor mobility, total distance traveled), conversation data (conversation duration and frequency), and location data (GPS, inferred buildings when indoors). While sleep duration had the strongest correlation with depression (pre ($r = -0.360$, $p = 0.025$) and post ($r = -0.382$, $p = 0.020$)), a strong negative association was found between conversation frequency during the day and depression (pre ($r = -0.403$, $p = 0.010$) and post ($r = -0.387$, $p = 0.016$)). Likewise, shorter conversation duration was an accurate predictor of a higher PHQ-9 score ($r = -0.328$, $p = 0.044$). The same held true for students that had fewer co-locations with other students ($r = -0.362$, $p = 0.025$).

2.7. Facial and Vocal Prosody

In the past decade, research has advanced in various applications of paralinguistic speech studies that detect information beyond the meaning of the words said (Sanchez et. al., 2011). Major depressive disorder can cause changes in one's neurophysiology alter speech production by influencing changes in vocal source and prosodics (Williamson et. al., 2013). Studies have shown that prosodic features such as speaking rate, voice energy, pitch and pause duration can be an effective way to detect the emotion state of an individual. Previously, clinicians have utilized verbal behaviors such as diminished prosody and monotonous sounding speech in support of the information provided by the patient for depression screening (Hall et al., 1995). In a similar context, primary health care settings can use speech processing methods to replace paper forms that are used in depression screening as a more effective and accurate procedure to assess depression. There has been sufficient progress made in researching effective methods of

computing mental health disorder such as depression detection using voice analysis. This section aims to review related work in this area.

In a depression recognition study based on dynamic facial and vocal expression features using partial least square regression, Meng et al, proposed a novel approach of using both visual and vocal cues to extract dynamic features from a video clip. They then described the emotional fluctuation of the two channels using three approaches, Motion History Histogram(MHH) for video extraction , adopting the spectra low-level descriptors (LLDs) and MFCC 11-16 that are included in the AVEC 2013 baseline audio feature set of the AVEC 2013 dataset for basic representation of vocal cues. They then used a regression algorithm called Partial Least Square (PLS) to map dynamic features to the depression state on each of the modalities. The performance was measured in The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) which were 9.02 and 11.4 respectively.

2.7.1. Dynamic Feature Extraction

The study "Depression Recognition based on Dynamic Facial and Vocal Expression Features" Meng et al, 2013, aimed at combining the dynamic feature of both the facial and oral expression. To extract video features they introduced Motion History Histogram (MHH) a human action recognition application to capture the movement of each pixel from the continuous image sequence within the face area. The edge orientation histogram (EOP) highlighted its temporal details as well as the Local Binary Patterns (LBP) where features are more concatenated for a better representation. MHH then records the change in gray scale for each pixel on the video. The extracted grayscale image from the video clip is what is known as the MHH feature M. The details of the dynamic features are then highlighted using EOP and LBP.

To compute the EOP of an image $f(u,v)$ the edges are first detected using the horizontal and vertical operators K_u and K_v as:

$$G_u(u, v) = K_u * f(u, v)$$
$$G_v(u, v) = K_v * f(u, v)$$

Using the avec 2013 baseline audio feature set such as the spectral low-level descriptors (LLDs) and MFCCs 1 -16, Meng et al, adopted the basic representation of vocal clues. The

AVEC 2013 dataset consists of 2268 baseline features set which is composed of 32 energy and spectral related and 42 functionals, 6 voicing related LLD and 32 functionals, 32 delta coefficients of the energy/spectral LDD and 19 functionals, 6 delta coefficient of the voice related LDD and 19 functionals and 10 voiced/unvoiced duration features. The LLD features are extracted from 25 ms and 60 ms overlapping windows which are then shifted at a rate of 10ms. Among these feature included pitch based LDD. The vocal changes in this study are considered to be discriminative between depression and other emotions using MHH, where instead of operating on each pixel in a video clip, components of the audio baseline features are used for dynamic modelling. For each component a sequence change is recorded with its corresponding histogram based on pattern variation. The final dynamic audio feature vector is therefore M times longer than the original one.

2.8. Detecting Depression Using Facial Coding

There is an abundance of papers that break down facial expressions into individual muscular actions, and correlate these muscular actions with a variety of moods and disorders. Human facial actions can be categorized in a variety of ways, but the one we will concern ourselves with is the academic standard: the Manual Facial Action Coding System (FACS). The FACS system was published in 1978, and it broke down facial expressions into a set of facial muscle movements that are called action units. The 1978 version had 64 distinct action units and more are added with every revision.

In their study on detecting depression from facial actions and vocal prosody, Cohn et. al. used two distinct methods to sense depression from facial data (2). The data is gathered by recording the face of an individual who suffers from major depressive disorder (MDD). The individual is asked three questions on their depressed mood, guilt, and suicidal thoughts. The responses to these questions comprise the dataset of the study.

The first method uses the aforementioned FACS coding to manually code action units frame by frame in a video of a participant. In this method, a FACS certified coder codes, or rather manually tags, the beginning (onset), the actual facial contraction (peak), and the ending (offset) times of an action unit. For each action unit, Cohn et. al. recorded 4 parameters: the proportion of the interview that each action unit occurs, mean duration, the ratio of the onset phase to total duration, and the ratio of onset to offset phase.

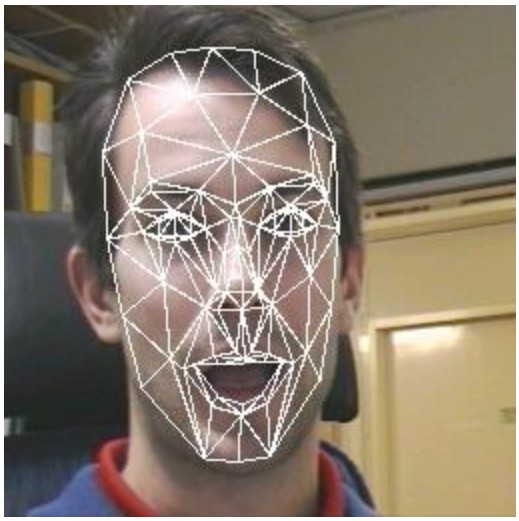

Figure 4: 2-D Triangular mesh created by AAM

The second method uses Active Appearance Modeling (AAM) to automatically create a 2D triangular mesh to overlay a participant's face. This 2D mesh is a collection of vertex locations after translation, rotation, and scale is removed. Normally, AAM is a person specific abstraction, but for the purposes of the paper a global model is created. These vertex locations correspond to a source image, from which the shape is aligned.

2.8.1. Shape of 2D Mesh:

$$2D\ Mesh\ =\ s0\ +\ \sum_{i=1}^{m} \sum s_i p_i$$

Where s is the base shape, and the summation sums m shape vectors with coefficients p which are shape parameters. The p values here are linear multipliers that essentially define the length of a shape vector. Differences in shape vectors represent changes in the actual object shape, such as muscle contractions.

For AAM, the data was converted to features by segmenting each interview into 10s intervals, then computing the mean, median and standard deviation of frame to frame differences in the p coefficients. On the sequence level, these statistics are combined by taking their mean, median, minimum, and maximum values. Thus for each eigenvector there are 12 statistics, and

10 main eigenvectors that correspond to certain muscle, resulting in 120 features per sequence. These features are fed into an SVM that utilizes a Gaussian kernel for classification.

The results for the FACS method pose some interesting insights into the problem of sensing depression from facial data, and might have some crossover applications with AAM. In FACS coding, the contraction of the buccinators muscle, which is a muscle used to tighten the corners of the lip, was positively correlated with depression. While it is not explicitly mentioned, the eigenvectors in the AAM likely picked up on this correlation to some extent, and any future implementation of this work could use this fact to increase their accuracy.

For both coding techniques, considerable accuracy in depression was achieved. With FACS coding, 79% accuracy was reached. This percentage is dwarfed by AAM, which resulted in 86% accuracy. It should be noted that within the context of our project, FACS coding posits an impractical way of sensing depression. Its full implementation requires manual coding, which does not fit in with our use case. The usage of AAMs would pose some problems as well, since the process isn't fully automatic. In the study, Cohn et. al. manually label %3 of key frames, and the remaining frames automatically align the 2D mesh using gradient-descent AAM fit. (1) If there was a way in which the AAM method could be utilized without the need for manual labeling, facial data could prove to be very useful in detecting depression.

2.9. Review of Modalities

Amongst all the modalities we surveyed, there seems to be no single golden bullet which presents perfect robustness in detecting depression. Every single modality presents promising results, thus the tradeoff must be evaluated considering the relative ease of providing a modality, or how likely the average person is to provide a certain modality. The analysis regarding other dimensions of these modalities are presented in further chapters.

3. Methodology

First we ran an exploratory "willingness to share data" study to determine what information patients feel comfortable giving to a member of medical staff. Then we developed an application capable of scraping the information from smartphones that we found patients would feasibly allow us to obtain. We used the application in conjunction with a survey to gather a large body of data and PHQ-9 scores to use to train machine learning systems. After we trained the machine learning systems, we gathered more data in order to validate the machine learning models learned from the data. Finally, we developed an Android application that could gather data from patient's phones and determine a value that represents the predicted severity of depression in the patient to present to their doctor.

3.1. Exploratory Study on Amazon Mechanical Turk

In the efforts to assess the public perception on sharing different kind of data from their mobile devices we conducted a survey. Our main goal of the exploratory study was to understand the different demographics and what types of information they are willing to share with a medical staff. The study also inquired about participants' willingness to perform tasks while being recorded by medical staff. The responses to the study were used to decide which modalities could feasibly be used to develop the final application and which modalities we abandoned because there is only a small chance that patients will be willing to give the necessary information. The full survey can be seen in Appendix B.

Our main method of collecting the data was through a survey, as we felt it was the most appropriate for collecting data from a large sample size of the general public. We employed Amazon Mechanical Turk to conduct the survey because it has the ability to reach a large number of people of a wide variety of demographics (Ipeirotis 2010). The data collected from Mechanical Turk was analyzed and used to draw conclusions on what information respondents were willing to share with medical staff. All questions asked were close ended questions where the participants would rank their willingness as one of five given options, so we used qualtrics to analyze the quantitative data.

To ensure validity of our findings in our survey, we included a general introduction of our research so that participants in the study understood the study's purpose and the importance

of answering the questions to the best of their abilities. We also gave the respondents a financial incentive so that they were willing to invest time to complete the survey. To protect the human subjects the entire survey was anonymous, a fact which all participants were be made aware of. The results of this study are discussed in depth in chapters 5.1 and 6.1, but since they drove the formation of the rest of the methodology I will briefly summarize the results. We found from the exploratory study that participants were most willing to share recordings of their voice and images of their face, and least willing to share their browser history and messages from chat apps such as GroupMe, WhatsApp, and Discord.

3.2. Development of a Data Scraping Application

Once we determined which types of data were feasible to obtain, we created an Android application capable of gathering each type of data from Android phones. The data gathered was used to train our machine learning systems so that they could predict depression levels. We first built the basic functionality of the application by building the functions for reading data from the phone. We then built simple web server and hosted it on a WPI virtual machine for receiving data from the application. Once we had an application capable of reading and sending data to our server, we put together the user interface components, focusing on making the application visually appealing and easy to understand and use. We then integrated our application and web server with the Twitter, Instagram, and Google API's and set up a page of the application for requesting users to sign in to their accounts with the three social media accounts listed.

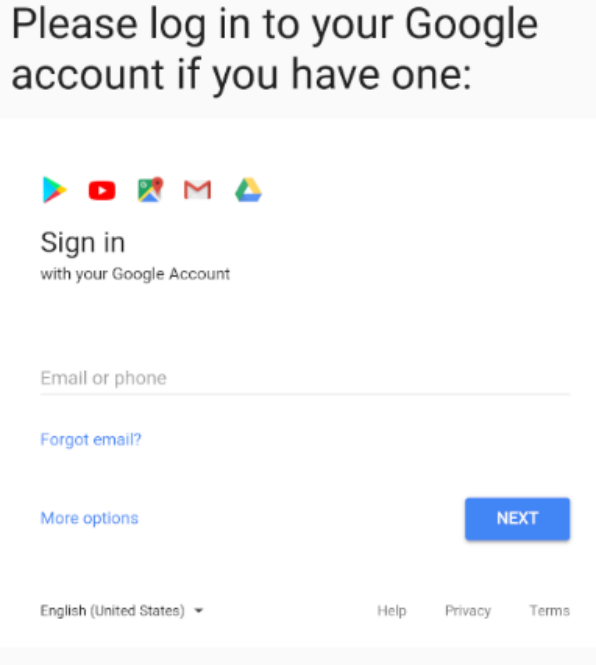

Figure 5 - Example login page for Google

Once the application was nearly finished and ready to be used in a survey, we gathered a small group of people to test the application experience on. We presented the participants with the small prompt they would receive if they were participating in the full survey, and then did not interfere while the participants worked through the application. In order to better understand each participant's experience, we also asked each participant to "think out loud," or verbalize their thought process. Based on the results of the test, we tweaked aspects of the user interface to make the flow of the application more understandable such as making the button for recording audio samples red before recording, and green after recording. Next, we ran a small study on the platform we planned to launch the full survey on, Amazon Mechanical Turk. We launched a survey that was the exact same as the full survey, except that we limited it to only 10 participants. The small version of the full survey allowed us to ensure that the application was working properly and that we were receiving the data.

3.3. Gathering Data

Once we had our scraping application fully functional, we began to gather data from actual users. We ran a Mechanical Turk survey that asked participants to install our application and fill out as much information as the participants felt comfortable providing. The PHQ-9 was

the depression measure we chose as the control variable in the study because it is one of the most commonly used methods of depression screening currently. Therefore, the application required all users to fill out the PHQ-9 which was provided in the application.

The application read and sent all data that the participant provided to our web server, which then sorted the data and stored the data. The data was stored according to a unique identification number assigned to each participant by the server upon opening the application. Therefore, we stored all data anonymously to protect participant's privacy. We decided to make the survey anonymous in order to make the participants feel safer and more comfortable giving data, which we believe increased trust and participation. However, since data was stored anonymously, we were unable to contact any individual whose PHQ-9 scores indicated severe depression in their PHQ-9 scores in order to provide them with resources that could help. If we had more time, it would have beneficial for us to create an automated way of presenting this information to participants who had a high depression score.

We stored the data on a WPI server in order to allow our machine learning algorithm to quickly and easily access the data, which significantly sped up computation. Once the data was gathered from users, we began to pre-process the data and prepare it for training with our machine learning algorithms. We then started pre-processing the data for use with our machine learning model.

The data collected included the user's calendar, contacts, call logs, text, and mobile files, voice recording, GPS locations over the past two weeks, Instagram posts, and Twitter posts. We requested users to give permission to the data on their mobile phone, but users could opt not to give access to some or all data. The full description of the application is presented in the Implementation chapter.

3.4. Machine Learning General Overview

Our sensing system utilizes machine learning methods to make sense of the large amounts of data that will be gathered. Our task is a multivariate learning task, the multivariate features being derived from the large number of modalities our project aims to use, and the label being the PHQ-9 score.

We see the limitation in using the PHQ-9 score as a way to generate depression labels, that is it does not present perfect sensitivity and selectivity, but by the virtue of requiring no

human input other than the interviewee's, we believe that this method allowed us to gather much more data than we would be able to by using a psychiatrist diagnosis for generating labels in our dataset, and still be of relative usefulness since we only allowed certain Amazon MTurk users who have successfully accomplished more than 50 tasks on the platform, and these users inputs could have been rejected by any of their employers on the Amazon MTurk platform.

The PHQ-9 score distribution of the training set can be seen in Figure 6. The distribution follows the occurrence of depression in the target population.

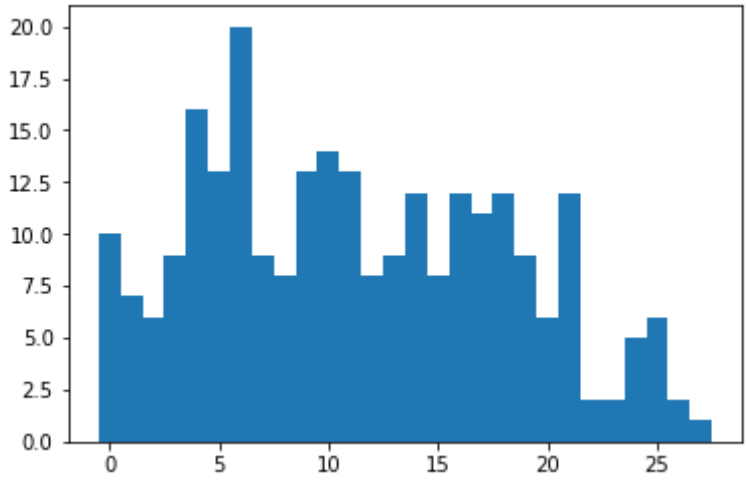

Figure 6: Depression score distribution

The machine learning techniques we use in our final application are varied. These techniques were chosen from many available types of classifiers and regressors that machine learning platforms offer. These classifiers differ by modality since we pick the best performing classifier for each modality. By modality, we mean a subset of data from a single source such as voice samples, location traces or Facebook. For example, the Instagram modality would present us with images of a certain hue, number of comments, likes etc. The numerical representation of these data that can be represented as a metric for every participant, and this metric would essentially be a feature.

We combine these learners to provide a summary score of depression, and we do this combination through using ensemble methods. Ensemble methods are methods that create multiple classifiers and combine them to produce improved results. Weighted averaging is an ensemble method in which the outputs of all classifiers are normalized and averaged (given that these individual classifiers all predict a continuous value for PHQ-9 results) using arbitrary weights at each classifier output to assign relative importance to that result.

Our individual learners that contribute to our ensemble method are trained using features whose creation we discuss below.

3.5. The Features Extracted

To perform machine learning, we needed to preprocess the data. The preprocessing step corresponds to all transformations applied to the raw data before it is used for training and testing. One essential step of preprocessing is feature extraction from the raw data, i.e., turning raw data into numerical features for use in training and testing. Below are details on the features we extract from our data for every modality.

3.5.1. GPS

This modality provides us with raw data in the form of a stream of GPS data for a given user. Previous studies we reviewed extracted two main types of features: various measures of movement and data on where the person is located. Our GPS data is in the form of a kml file, which is stores geographic modeling information in XML format, and contains place marks that contain the information of a user's location such as name, description, timestamps, point coordinates among others.

Since the nature of our data allows for the extraction of literature backed features, we extracted features belonging to the two categories mentioned above. These features include the number of different places visited, the maximum distance between two locations, the total distance covered, and how much a user perform a certain activity such as running, walking or driving. Below are the formulas of how from the Trajectories of Depression study; Lucas et al. calculate the above features and the rationale of why those features are important in monitoring depressive states by means of smartphone mobility traces. It is important to note that these formulas were slightly modified in our study to fit our specific dataset.

1) Number of different places visited: Denoting $N_{diff}(t1,t2)$ the number of different places visited between time intervals t_1, t_2 in the placemark. The different places are coordinate points pair of the longitude and latitude c_i and c_{i+1}. The formula is given as follows:-

$$N_{diff}(t_1, t_2) = \sum_{i=1}^{N(t_1,t_2)} count(c_i, c_{i+1})$$

2) The total distance covered: Denoted as $D_T(t_1, t_2)$, we will define this distance as follows:-

$$D_T(t_1, t_2) = \sum_{i=1}^{N(t_1,t_2)-1} d(C_i, C_{i+1}),$$

where $d(C_i, C_{i+1})$ is the distance between the two latitude-longitude pairs $c_i \; and \; c_{i+1}$.

3) The maximum distance between two location: Denoted as $\mathrm{D_M}(\mathrm{t_1}, \mathrm{t_2})$, it represents the maximum span of the area covered between time intervals t_1, t_2 .

$$\mathrm{D_M}(\mathrm{t_1}, \mathrm{t_2}) = \max_{i,j \in \{1,\ldots,\mathrm{N}(\mathrm{t_1},\mathrm{t_2})\}} d(C_i, C_{i+1}).$$

4) Number of times an activity occurs: Denoting $N_{act}(t_1, t_2)$as the number of times a user performs a specific activity between a given time interval t_1, t_2 in the placemarks.

$$N_{act}(t_1, t_2) = \sum_{i=1}^{N(t_1,t_2)} count(n_{active}),$$

where n_{active} is the name of activity in the a specific placemark.

In regards to the rationale behind the extraction of the above features; these specific feature relate more to how active a user is. According to a study conducted by the biological psychiatry review board; title "Physical activity, exercise, depression and anxiety disorders. It consistently associated high reported level of habitual physical activity with better mental health and a correlation of habitual high mobility traces with low depression rates(Andreas, 2008). The above feature extract information that indicates whether a user has high or low mobility traces and this would help in depression sensing .

3.5.2. Social Media

Our social media data includes twitter and Instagram data. For this modality, we extracted many features relating to engagement, sociability, and more. These features are explained below.

3.5.2.1. Twitter

From raw data we gather the features of average number of posts, replies, and retweets every day. This provides for an engagement metric. There is also data related to how many people are following/friends with a subject, and how many people follow each subject. This provides for an ego-centricity measure. For our project, we decided not to utilize twitter text data, since the response rate for Twitter was less than one thirds.

The features we used were like frequency, comment frequency, retweet frequency and liked frequency.

$$Like\ Frequency = \frac{aggregate\ number\ of\ likes\ on\ every\ tweet\ subject\ has\ posted\ in\ past\ 2\ weeks}{14}$$

$$Comment\ Frequency = \frac{aggregate\ number\ of\ comments\ on\ every\ tweet\ subject\ has\ posted\ in\ past\ 2\ weeks}{14}$$

$$Retweet\ Frequency = \frac{aggregate\ number\ of\ retweets\ on\ every\ tweet\ subject\ has\ posted\ in\ past\ 2\ weeks}{14}$$

$$Liked\ Frequency = \frac{aggregate\ number\ of\ likes\ the\ subject\ has\ given\ in\ past\ 2\ weeks}{14}$$

3.5.2.2. Instagram

The Instagram modality is interesting since it introduces machine vision into our project. This modality includes features from both social media usage, that is the comments and likes on a picture and number of followers and number of people followed, and visual data, such as the hue and saturation values of photos (HSV), filter information, or how many faces are in a posted picture.

To be precise, the features we used were follower count, following count, an 8 element vector that counts the usage of the Amaro, Crema, Hefe, Inkwell, Rise, Valencia, Willow, X-Pro II filters (these usage of these filters either positively or negatively correlated with depression) (21), like frequency, comment frequency, post frequency, one final pixel wise average of Hue, Saturation and Value for every picture posted, and number of faces on average for each post.

$$Filter\ Vector = [\frac{number\ of\ times\ Amaro\ was\ used\ in\ past\ 2\ weeks}{total\ number\ of\ photos\ posted\ in\ past\ 2\ weeks}, \frac{number\ of\ times\ Crema\ was\ used\ in\ past\ 2\ weeks}{total\ number\ of\ photos\ posted\ in\ past\ 2\ weeks}$$

$$, \frac{number\ of\ times\ Hefe\ was\ used\ in\ past\ 2\ weeks}{total\ number\ of\ photos\ posted\ in\ past\ 2\ weeks}, \frac{number\ of\ times\ Inkwell\ was\ used\ in\ past\ 2\ weeks}{total\ number\ of\ photos\ posted\ in\ past\ 2\ weeks}, \frac{number\ of\ times\ Rise\ was\ used\ in\ past\ 2\ weeks}{total\ number\ of\ photos\ posted\ in\ past\ 2\ weeks}$$

$$, \frac{number\ of\ times\ Valencia\ was\ used\ in\ past\ 2\ weeks}{total\ number\ of\ photos\ posted\ in\ past\ 2\ weeks}, \frac{number\ of\ times\ Willow\ was\ used\ in\ past\ 2\ weeks}{total\ number\ of\ photos\ posted\ in\ past\ 2\ weeks}, \frac{number\ of\ times\ X-Pro\ II\ was\ used\ in\ past\ 2\ weeks}{total\ number\ of\ photos\ posted\ in\ past\ 2\ weeks}]$$

$$Like\ Frequency = \frac{aggregate\ number\ of\ likes\ on\ every\ Instagram\ photo\ subject\ has\ posted\ in\ past\ 2\ weeks}{14}$$

$$Comment\ Frequency = \frac{aggregate\ number\ of\ comments\ on\ every\ Instagram\ photo\ subject\ has\ posted\ in\ past\ 2\ weeks}{14}$$

$$HSV\ Vector = [\frac{sum\ of\ all\ pixel\ hue\ values\ on\ every\ Instagram\ photo\ subject\ has\ posted\ in\ past\ 2\ weeks}{count\ of\ all\ pixels\ in\ every\ Instagram\ photo\ subject\ has\ posted\ in\ past\ 2\ weeks},$$

$$, \frac{sum\ of\ all\ pixel\ saturation\ values\ on\ every\ Instagram\ photo\ subject\ has\ posted\ in\ past\ 2\ weeks}{count\ of\ all\ pixels\ in\ every\ Instagram\ photo\ subject\ has\ posted\ in\ past\ 2\ weeks}$$

$$\frac{sum\ of\ all\ pixel\ value\ values\ on\ every\ Instagram\ photo\ subject\ has\ posted\ in\ past\ 2\ weeks}{count\ of\ all\ pixels\ in\ every\ Instagram\ photo\ subject\ has\ posted\ in\ past\ 2\ weeks}]$$

$$Average\ Number\ of\ Faces = \frac{sum\ of\ number\ of\ faces\ in\ each\ Instagram\ photo\ subject\ has\ posted\ in\ past\ 2\ weeks}{total\ number\ of\ photos\ posted\ in\ past\ 2\ weeks}$$

3.5.3. Text Data

For text data, in addition to the text found in tweets and Instagram posts, we also have every participant's cellular texts. Text data can be featurized using a variety of methods. In our paper, we concentrate on sentiment analysis, part of speech tagging and the frequency of texting.

Our features feature a 14 day 5-day moving average of sentiment scores, 14 day 5-day moving average of texting frequency, and a 45 column vector that has for each element the frequency of usage of a particular part of speech tag using the Penn Treebank.

We use the list of 45 parts of speech the Penn Treebank provides. The Penn Treebank is a collection of many syntactic trees that have been generated over 8 years between 1989 and 1996. The methodology they use is of particular interest to us, since this methodology is a way to biject any text into its constituent parts of speech (Taylor, 2003). These 45 parts of speeches are linguistic descriptions of all words that perfectly bisect the English language in 45 types of words. While there is no study that has observed a correlation between depression and part of

speech tags, there is suggestive evidence that part of speech tags are a robust way to feature text, and that they might correlate with certain moods and personality types.

$$Sentiment\ Score\ Moving\ Average = \sum_{i=1}^{14} \frac{sum\ of\ every\ words\ sentiment\ score\ in\ every\ text\ sent\ in\ the\ 5\ days\ subsequent\ to\ ith\ day\ of\ past\ 2\ weeks}{count\ of\ every\ word\ in\ every\ text\ sent\ in\ the\ 5\ days\ subsequent\ to\ the\ ith\ day\ of\ past\ 2\ weeks}$$

$$Texting\ Frequency\ Moving\ Average = \sum_{i=1}^{14} \frac{count\ of\ every\ text\ sent\ in\ the\ 5\ days\ subsequent\ to\ ith\ day\ of\ past\ 2\ weeks}{count\ of\ every\ text\ sent\ in\ the\ 5\ days\ subsequent\ to\ the\ ith\ day\ of\ past\ 2\ weeks}$$

$$POS\ Tag\ Vector\ (45\ long) = [\frac{count\ of\ coordinating\ conjunctions\ used\ in\ every\ text\ send\ in\ past\ 2\ weeks}{count\ of\ every\ word\ in\ every\ text\ sent\ in\ past\ 2\ weeks}, \ldots$$
$$, \frac{count\ of\ prepositions\ or\ subordinating\ conjuntions\ used\ in\ every\ text\ send\ in\ past\ 2\ weeks}{count\ of\ every\ word\ in\ every\ text\ sent\ in\ past\ 2\ weeks}]$$

3.5.4. Call Data

For call data, we only have metadata for phone calls. This includes, amongst other things, the time of the call.

For call data we employ another 14 day 5-day moving average of call frequency.

$$Calling\ Frequency\ Moving\ Average = \sum_{i=1}^{14} \frac{count\ of\ every\ call\ made\ in\ the\ 5\ days\ subsequent\ to\ ith\ day\ of\ past\ 2\ weeks}{count\ of\ every\ call\ made\ in\ the\ 5\ days\ subsequent\ to\ the\ ith\ day\ of\ past\ 2\ weeks}$$

3.5.5. Audio

The audio modality presented us with an interesting challenge; all the work that was done prior to this study that is on correlating voice with depression used hand coded, ~30 minute recordings of audio. Our use case restricts us to detecting depression on the spot, so we decided to gather 10 seconds of audio from every person. Since inter-variability in sound data would prove a challenge, we made every participant speak the standard phrase “The quick brown fox jumps over the lazy dog” into their phones.

We use openSMILE, a feature extraction tool to extract around 1600 features from these sound files. openSMILE is an acronym for “The Munich open Speech and Music Interpretation by Large Space Extraction”. It is a toolkit to extract audio-signal features from any sound file, and is implemented in C++ (Florian, 2010).

We use the emobase 2010 configuration file that comes prepackaged with openSMILE

3.5.6. Contacts

For contacts we use the number of contacts the subject has on their phone as a feature.

$$Number\ of\ Contacts\ = \sum contacts\ on\ subjects\ phone$$

3.6. Machine Learning Architecture

Our machine learning workflow can be broken down into four stages:

1. Restructuring the database
2. Extracting features from the data
3. Preprocessing the data
4. Training classifiers

The step of restructuring the database was essential since the data was stored in an unorganized manner. There were multiple entries for the same modality and person that needed to be merged together. For such entries, the entries were merged, and all the data was saved on disk as a byte stream (a python pickle), one byte stream for every modality and user. An example of this process is illustrated in Fig. 2.

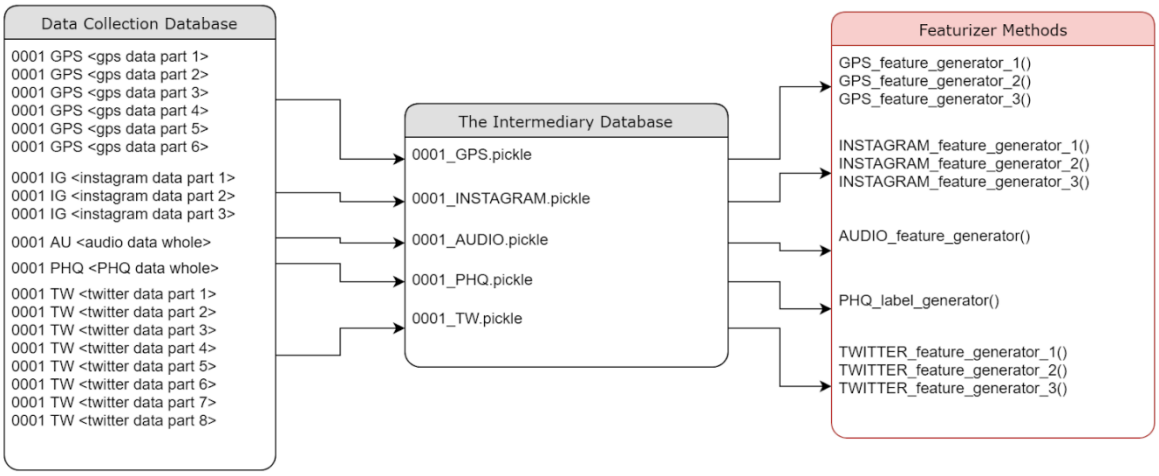

Figure 7: Intermediary database

For the second step of extracting features from the data, we wrote functions that took in a certain users certain datatype (eg. audio for user 4200), and output a feature or a number of features for that particular person and datatype. We call this a feature vector, and one of these

variable size feature vectors exist for every combination of participant and modality. This process is illustrated in Fig. 3.

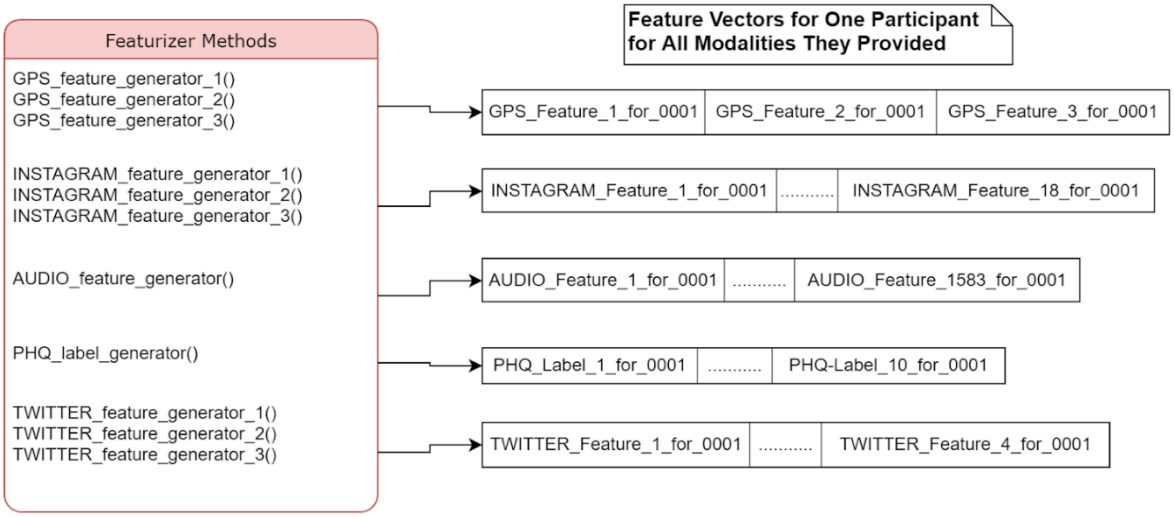

Figure 8: Creation of feature vectors

We then ran these functions to generate features for each user, and glue the resulting column vectors vertically to form a matrix of size (number of participants in study, number of features + label). This matrix of features allows us to have the right dimension and type of format for use with scikit-learn learners. The creation of one row of this matrix is illustrated in Figure 9.

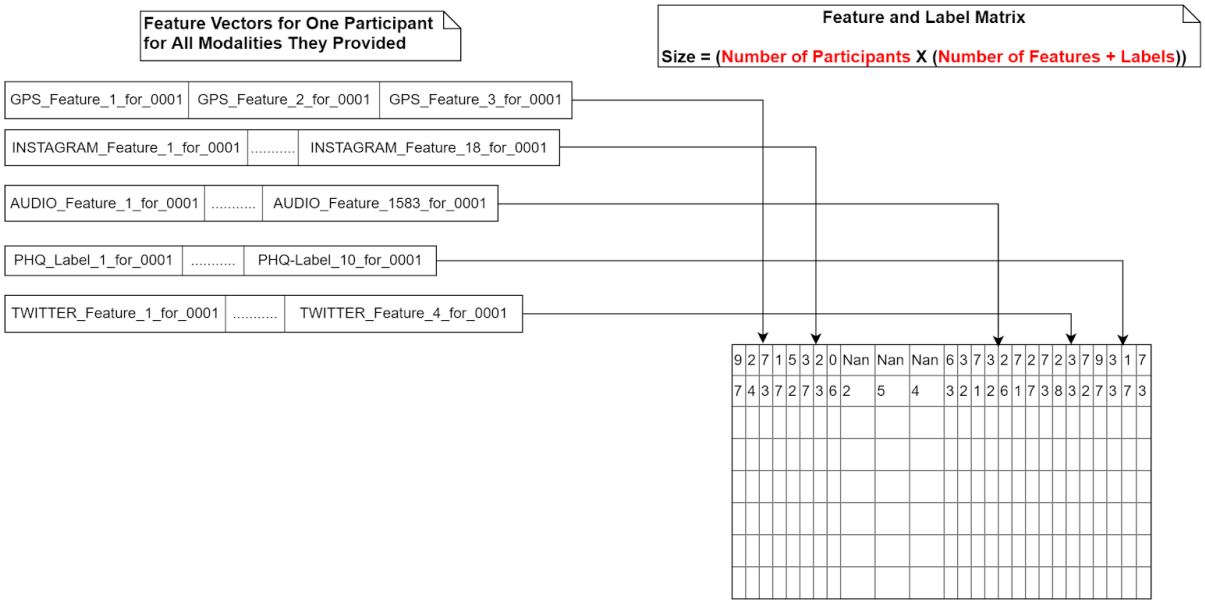

Figure 9: Creation of deature matrix

To avoid data snooping and to scientifically see if the classifiers that were trained do generalize over their intended population, 85% percent of the data was designated as the training set, while the remaining 15% of data was put away as the testing set. All the development concerning classifiers was done on this training set, using k-fold cross validation.

In preprocessing the data, we followed basic, statistically sensible steps that would prepare our data to be inputted into classifiers. To be specific, we first divide the above matrix into 7 different submatrices that correspond with all the different modalities we have. The features are contiguous in a modality context, so these 7 submatrices all contain the features for every user for a given modality. Then we shuffle all of these submatrices and normalize the data so it has a Gaussian distribution, zero mean and unit variance.

In an earlier iteration of this project, the big matrix mentioned above wasn't divided into submatrices, so missing values were replaced with the mean values. In our current implementation, we discard entries with missing data for every modality.

The last step was training classifiers. For this step, many classifiers and regressors that scikitlearn provides were used and compared on these individual sub-datasets. These learners include linear regression, random forest classifier, random forest regressor, SVM for classification, SVM for regression, a basic multi-layer perceptron with one hidden layer (neural network), logistic regression, naive bayes and knn. The training of these learners were done by training them on the training set with k-fold cross validation, and hyperparameter search was used to find the best parameters possible for these learners. After hyperparameter optimization, we use the bagging ensemble method in the hopes of reducing overfitting of our learners. This process is illustrated in Fig. 5.

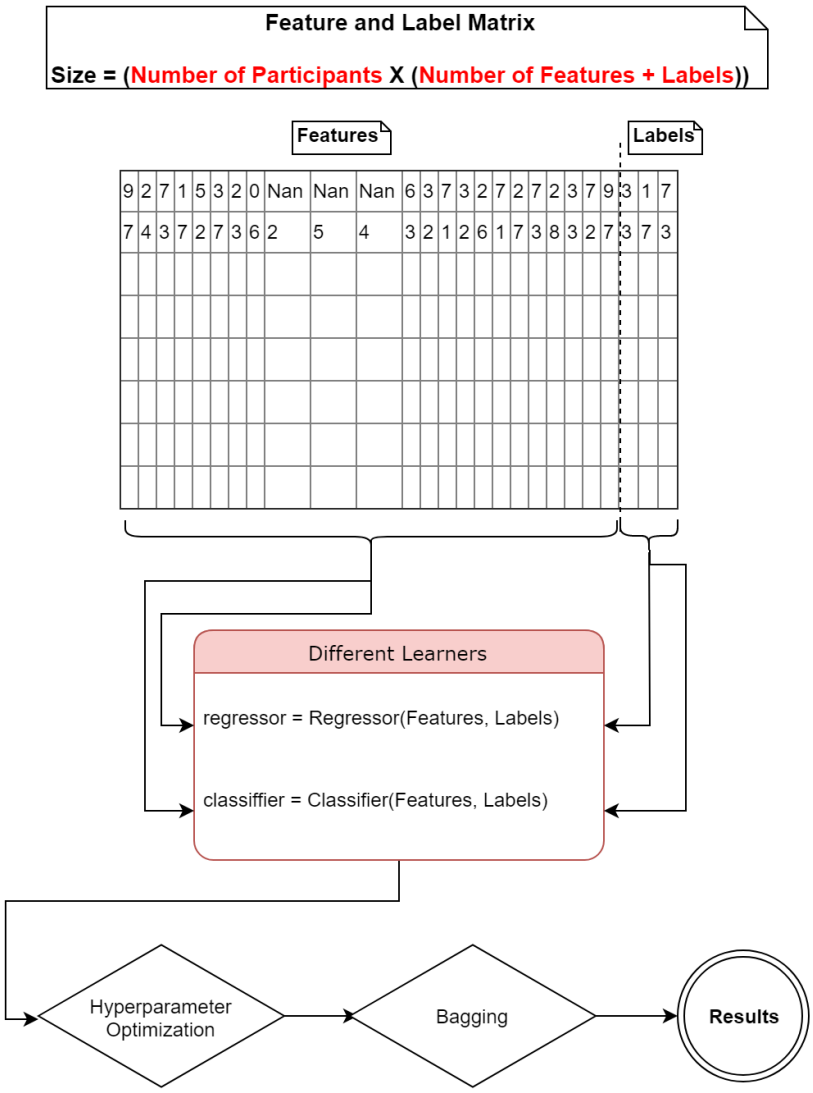

Figure 10: Training of learners

For regression tasks, we trained the classifiers by treating PHQ-9 scores as continuous values. We report the root mean square error (RMSE) at the end. The RMSE is a measure of the differences between values that are predicted by a model and the true values the model tries to predict. It is found by taking the square root of mean squared error (MSE). The mathematical definition for RMSE is provided below:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Figure 11: RMSE equation

In the above equation Y hat is a vector of n predictions. Y is the vector of observed values of which the predictions are made for.

Our label exists as a PHQ-9 aggregate score, where it exists as an integer ranging from 0 to 27. We create binary labels using this continuous label at certain cutoffs. Pfizer, the creator of the PHQ-9 depression questionnaire has identified the 10, 15 and 20 cutoffs of this aggregate score to correspond with moderate, moderately severe and severe depression. Therefore we create three binary labels using the cutoffs 10, 15 and 20, to make it so that our learners can be trained to recognize these particular severities of depression. These binary labels represent different severities of depression, for which a 1 indicates that the PHQ-9 score is higher than the cutoff and 0 indicates the PHQ-9 score is lower than the cutoff. It should be noted that we balance the datasets for the aforementioned PHQ-9 cutoffs. For these classifiers we report precision, recall and support.

We ran hyper parameter optimization algorithms on all of these classifiers, and compared these results with other hyper parameter optimized classifiers for which we reduced the feature set. The feature reduction techniques used were picking the top performing features in a random forest classifier, and stability selection, which is the application of l1 regularization (lasso) to different subsets the dataset many times, until some features tend to express themselves as important for classification, and some features do not. The reason behind some features losing importance, or being labeled by a stability selector as unimportant, is since l1 regularization constricts the feature space into a diamond shape around the origin(see Fig. 1.), the local minima often shows itself at a point on either axes, therefore nullifying the predictive power of a feature.

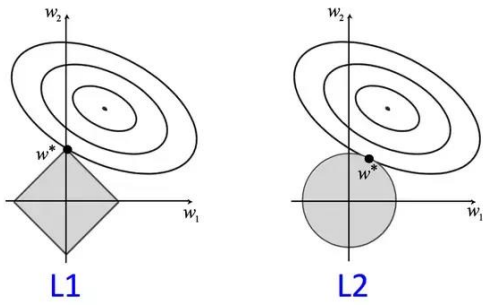

Figure 12: L2 Regularization

After hyper parameter optimization and feature reduction, the best performing classifiers were determined by looking at their confusion matrix, recall and precision.

3.7. Gathering Test Data

We gathered test data in order to test the performance of our application. We gathered participants' social media account handles and the permission to gather their posts, participant's GPS data for as far back as possible, and participants' phone information such as call logs, health app data, texts, and app usage. Alongside this information participants' were also asked to complete the PHQ-9 questionnaire to get a baseline depression rating to compare the results of the application to.

The test data was mainly gathered through online surveys. First, the participants were asked complete the PHQ-9. If the participant declines to answer the PHQ-9, the survey ended because without the baseline PHQ data to compare predictions to their data is not useful. Then, the survey simply had each participant download our application so that the application could send all their phone data to a server which saved the participant from having to fill out a lengthy survey. Finally, the user was asked to give their social media account handles and give permission to gather their posts on the corresponding sites. At any step of the application the participant could decline to give a piece of information. This mimics the real life cases where the application will be used where patients have complete control over how much information they give medical professionals. In fact, partial data sets allowed us to see how well the application performed with less information, as well as look for trends in which demographics typically do not feel comfortable handing over certain types of data.

The surveys were sent out over Amazon Mechanical Turk with a financial incentive to complete the survey, as well as an additional financial incentive for giving each type of data (see section 4.1 for more details on the incentives). We chose Mechanical Turk because a wide variety of demographics use Mechanical Turk (Ipeirotis 2010). We wanted to avoid sampling from any specific demographic because in practice the app will be used on all types of people from the general public. Mechanical Turk also allows us to provide a financial incentive to complete the survey which could increase the number of results we get from the surveys. We aimed for approximately 400 responses.

3.8. The Final Application

With a complete machine learning algorithm that could predict the level of severity of depression in a patient, we began to build a final production build of our Android application for use in hospitals. The previous build was made for use in a survey, so we first removed all parts of the application that explained how to complete the survey portions of the application, such as submitting their code to Mechanical Turk. We also removed the PHQ-9 from the application, because it was present to be used as a control variable, which we no longer needed.

Since the final version is intended for a more widespread and professional use, we spent the rest of our time developing the user interface of the application to make it look smooth, friendly, and professional. We contacted user experience experts at WPI to ask what changes they would like to see in the application. We then tested the application on a few participants and once again asked to vocalize their thoughts as we observed them interacting with the application. After making alteration based on the results of the small tests, we had a final android application capable of estimating a patient's level of depression.

4. Implementation

In order to develop the application, we had to create several systems. The systems we created included an android application for gathering data from users, a server for collecting and parsing the data, a database for storing the data, and machine learning systems for creating a model that can predict the levels of depression in a participant. The overall flow of the systems we created are summarized in figure 13.

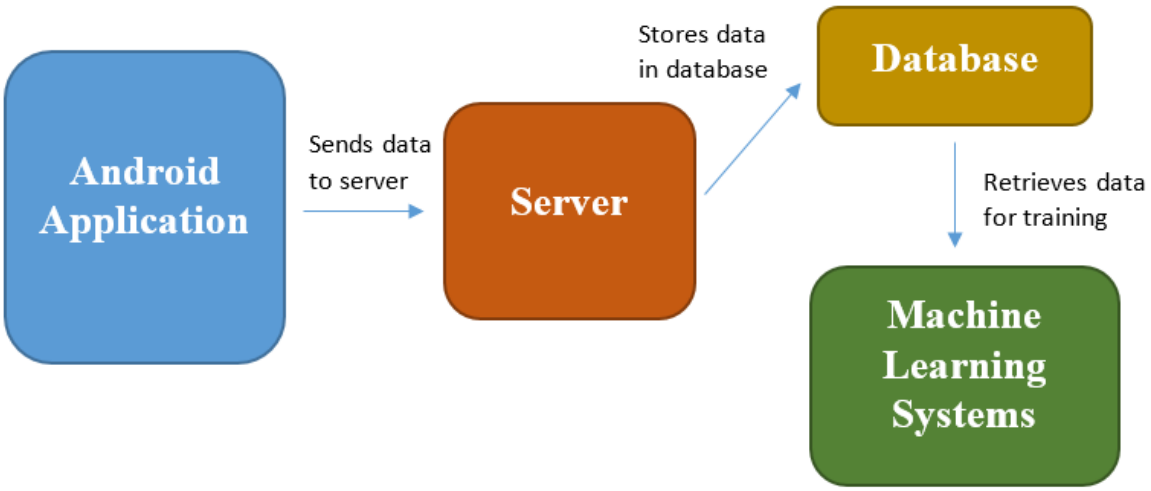

Figure 13 - The overview of our systems and how they connect

4.1. Android Application

In order to complete the study, all participants were required to download an Android Application through which they would complete the study. The application we designed consisted of a series of screens where the applicant was asked to complete certain tasks. Each task relates to one of the modalities we decided to use based on the exploratory survey results (see section 6.1). To summarize the findings of the exploratory study, participants were most willing to share voice recordings and images of their face, least willing to share browser history and data from text chat apps like GroupMe and WhatsApp, and about split for every other data type. Therefore we decided to use the text, call logs, calendar, system files, contacts, Twitter, Instagram, GPS, and voice recording modalities based on the willingness of participants in the

study and the performance metrics of the modalities in other works discussed in the Literature Review (Section 2).

Every screen has a next screen button at the bottom, and the applicant can return to the previous screen at any time with the back button on their phone. The top right of the application also keeps a running total of how much money the applicant would make for completing the study. The total started at $0.40, which was the base pay for the study and increased by $0.30 for giving Google GPS data, and by $0.10 each for completing a voice recording, giving a Twitter username, and logging in to Instagram.

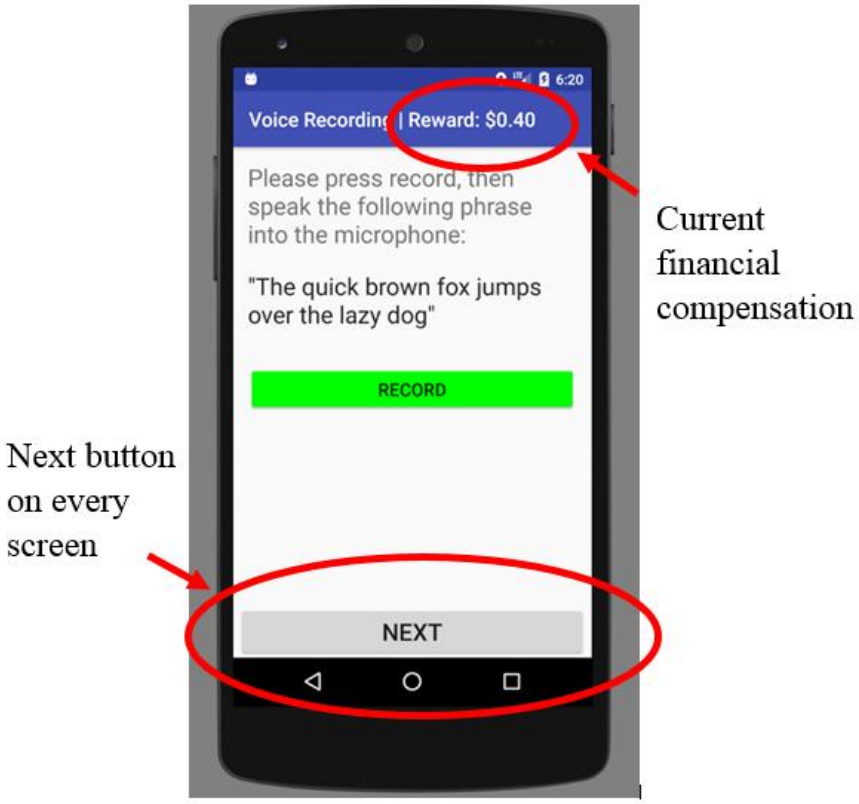

Figure 14 - Locations of “Reward” and “Next” button screen elements

4.1.1. Flow

The overall Android Application moves the five major steps. The five steps are as follow:

1. **Request Android permissions and send data that require permission** - The very first step was to ask for Android permissions so that the application could immediately start sending data in the background. This gave the application the most time for sending data possible, which was important because the amount of data to send, if large, could potentially take longer than the applicant would be willing to wait on a single screen for.
2. **Fill out the PHQ-9** - Applicants would then be required to fill out the PHQ-9 because it was the only step that required it to be filled out completely.
3. **Record a Voice Sample** - According to our exploratory study, participants were most willing to give voice recordings. Therefore, we then requested voice data so as to start with the easiest requests to keep the applicants willing to participate for as long as possible.
4. **Sign in to Google, Twitter, and Instagram** - According to our exploratory study, participants were less willing to give GPS data and social media posts. Therefore, we requested that applicants sign in to Google, Twitter, and Instagram last in case the applicants became offended by the requests or decided to stop filling out the study.
5. **Thank You and Code Screen** - Finally, the application gave applicants the code they needed to prove they completed the study.

4.1.2. Screen 1 – Outline Screen

The first screen consists of a brief outline of the entire application so that applicants knew what they would be asked to complete and know how much they had left. The outline looked like the overview of the application given in the previous section (section 1.1.1). Below the outline, there was a disclaimer that reminded applicants that they have the choice to refrain from giving any information they did not feel comfortable giving. As seen in Figure 15, the disclaimer read:

"IMPORTANT: Please remember that you are not required to fill out any section of this study. However, all information you do give will be stored anonymously on a secure server and will not be tied to you. The more information you give the more effective we can make this app at detecting depression early which could potentially save lives, so please answer as much as you can!"

While the applicants read the page, the application sends an http request to our server which responds with a unique ID number. The unique ID number is then used to identify all data

gathered throughout the entire application to keep all data anonymous. If the application fails to receive an ID, the application would redirect to the “Connection Failed” page.

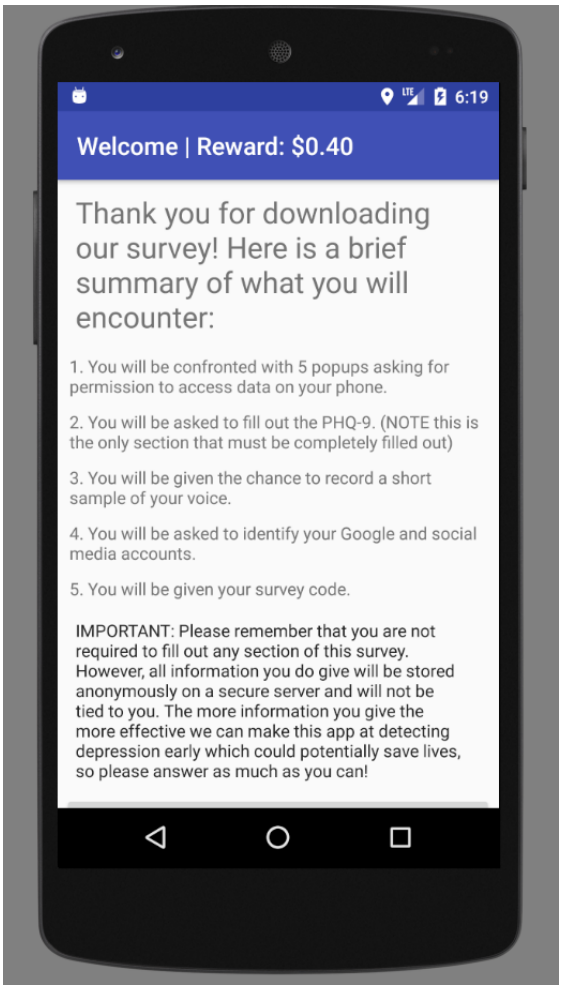

Figure 15 - Screen 1 of the data gathering application

4.1.2.1. Connection Failed Page

The “Connection Failed” page simply displays text explaining that no internet connection with the server was obtained and that that an internet connection must be established in order to continue. It is important to obtain a code before continuing because without the code there is no way to connect the data that is sent with the control variable. When the applicant presses the next button, it will retry the connection. If the connection fails once again, the applicant remains on the Connection Failed page. If the application successfully obtains an ID, the application continues to Screen 2.

4.1.3. Screen 2 – PHQ-9 Screen

As seen in Figure 16, the first thing that applicants saw on screen 2 is a popup window with five screens. Each screen asked the applicant to give an Android permission. The permissions requested were the permissions to read the calendar, read the phone's contacts, read the phone's call logs, read the phone's saved text messages, and to read and write to the phone's storage. The applicant can refuse or accept each permission individually. Once all permissions have been addressed, the popup window disappears and the application returns to Screen 2. For each permission that was granted, the application began sending the corresponding data to our server in the background while the applicant complete the remaining portion of the application. Each permission granted starts up one background thread that sends the corresponding data so that all data types are sending simultaneously to send data faster. The faster that the application could send the data the more data would be collected before the applicant closes the application.

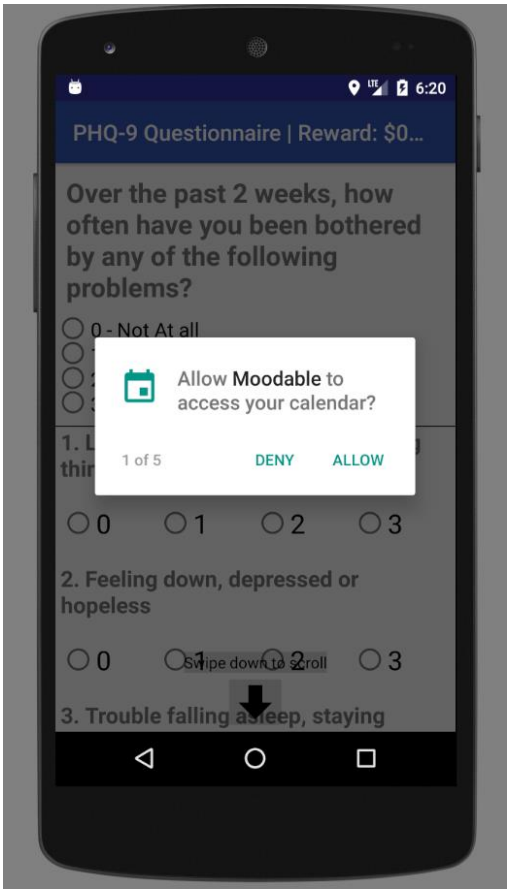

Figure 16 - The permission requests on the second page of the data gathering application.

Screen2 has the applicants fill out the PHQ-9, which is paper based depression test that is currently used. The PHQ-9 is used as our control variable in this study. As shown in Figure 17, Screen 2 first explains the process for filling out the PHQ-9, and then uses a scrolling screen to give all 9 questions. All questions on the PHQ-9 are required to be answered before the applicant can continue to Screen 3, because the PHQ-9 is the only required part of the application.

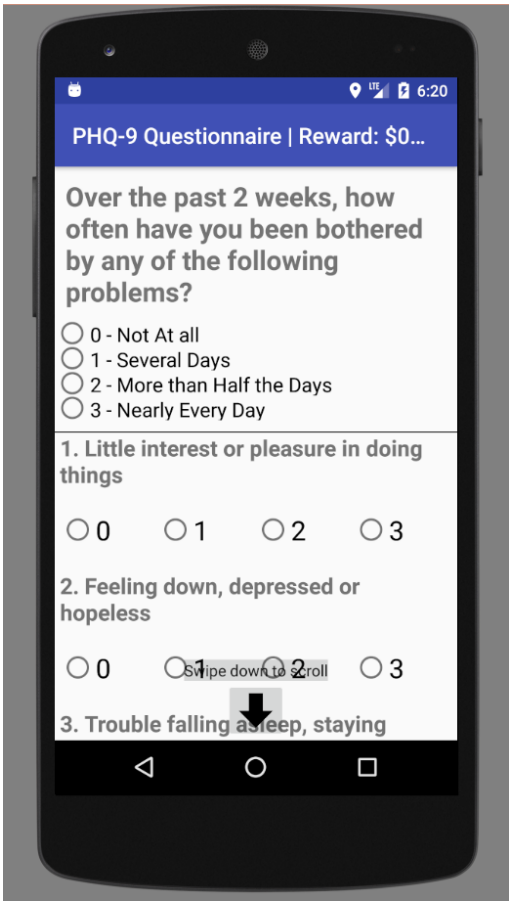

Figure 17 - Screen 2 where applicants fill out the PHQ-9

4.1.4. Screen 3 – Voice Recording Screen

The third screen asked applicants to record a short recording of their voice (see Figure 18). Another small popup window appears when first visiting this page asking for more Android permissions. The permissions requested are permission to record audio through the phone's microphone, and permission to read and write to storage (if not already given). The top of the page reads: "Please press record, then speak the following phrase into the microphone: The quick

brown fox jumps over the lazy dog." The center of page has a big button that labeled “Record” that when pressed starts recording through the phone’s microphone. Upon pressing, the “Record” button becomes a “Stop” button, which the applicant presses after reading the phrase. Once a recording is successfully obtained, text pops up thanking the applicant for recording so that they are aware that the recording saved. The applicant was free to press the Next button at any point without recording anything.

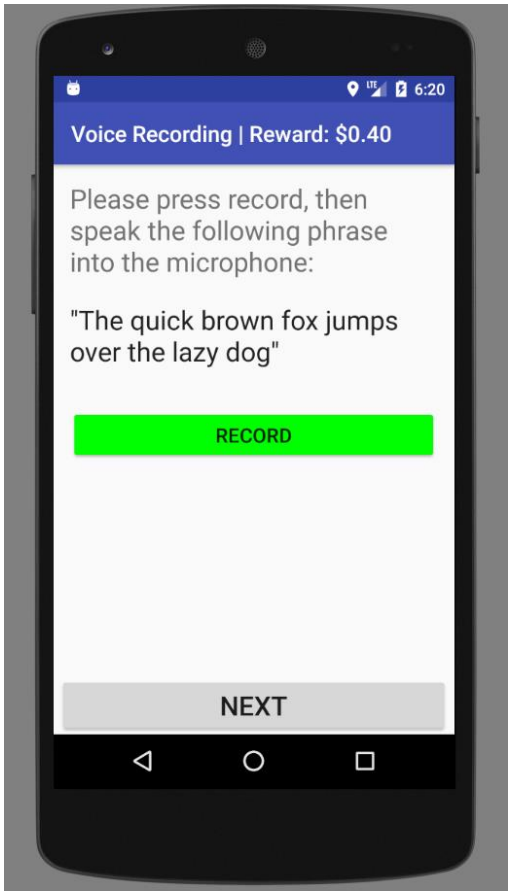

Figure 18 - The voice recording page of the data gathering application.

4.1.5. Screen 4 – Social Media Screen

The fourth screen asked applicants to sign in to three accounts: Twitter, Instagram, and Google. The page explains that by signing in to each account is also giving us permission to gather posts from their account, which can be seen in Figure 19. The top of the screen contains a small window containing a web page with the google sign in page open. If the applicant signed in to their google account, the web page would then open up 14 urls to download pages for the last 14 days of GPS data if any GPS data is saved in their Google account. Google gathers this

data from various applications that people have installed, and make it available to any of their users as long as they have signed in to their account. Once the Google GPS data was downloaded (in .kml format) it was sent to our server to be collected. Second, applicants were asked to provide their Twitter username. If applicants provided a username, the username was sent to our server where it would be used to retrieve their last 200 tweets. At the bottom of the screen is another small web page with the Instagram login page open. If the applicant signed in to Instagram, the web page would retrieve an access token from the Instagram API and deliver it to our server. The server would then use the access token to gather the Instagram account's account information and posts. Once again, the applicant could simply skip the entire page without giving any of the account's information.

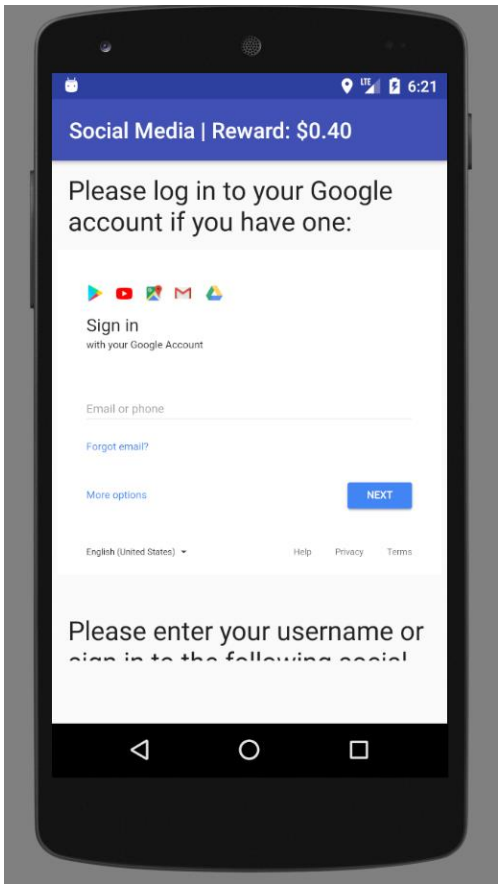

Figure 19- The social media page of the data gathering application.

4.1.6. Screen 5 – Survey Code Screen

The fifth screen is the last screen of the application, which is shown in Figure 20. When the applicant reaches the Screen 5, any data being sent in the background stops sending, and a

“sending complete” message is sent to the server. Screen 5 simply displays a message thanking the applicant for completing the study, and presents the applicant with a confirmation code to give Amazon Mechanical Turk proving that they completed the study. The code also contained information that we could use to see how much information the applicant gave so that we could properly pay the applicant for their responses.

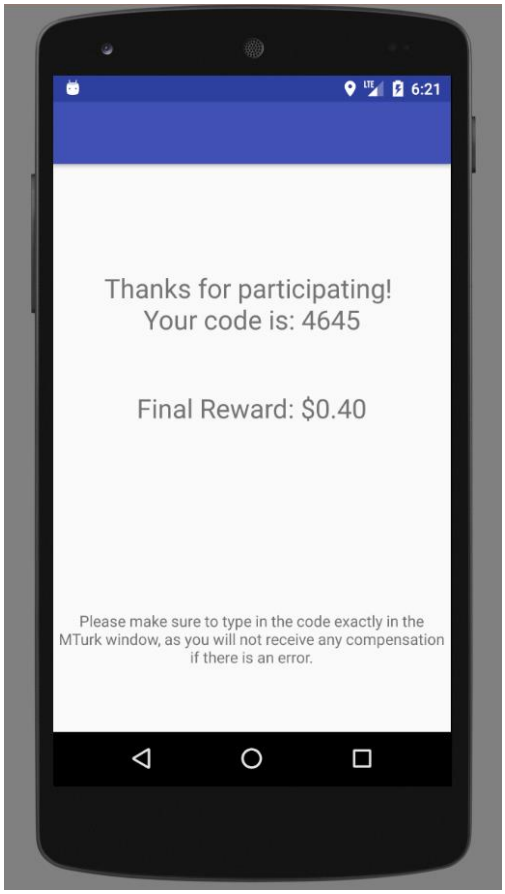

Figure 20 - The last page of the data gathering application where users are given the survey code.

4.2. The Server

The second piece of software developed for the data gathering study was a Nodejs server for receiving and storing data sent from the Android application where the data was converted into JSON objects and collected into groups of data for each participant. When the server is running, it listens for messages sent through HTTP. Each message sent to the server must contain the ID number of the applicant who sent the data, the type of data being sent, and data to be stored as a string. When the server receives a message, it inserts it into a SQLite database.

However, for the Android permission data that is being sent continuously, it collects all the messages in memory and holds it there until a “sending finished” message is received from the same ID. When the “sending finished” message is received the server will then insert all messages saved from the ID into the database at once, which speeds up the insertion process. In order to test the server, a simple web page was set up for us to view the data stored in the database. Once the study went up, we changed the web page to instead only display how much data was stored instead of displaying the actual data so as to preserve the privacy of the data.

4.2.1. The Database

The database is a simple SQLite database that uses a simple database file to store data. SQLite was used for the ease of learning so that we would have more time to develop other aspects of the software, but another database system such as PostgreSQL or MySQL would be a more robust option. The database consisted of two tables: one for User IDs, and one for storing data. The user ID table stored each user ID that has data stored along with the timestamp of when the applicant with the ID first sent data. The user ID table allowed for fast and easy lookup of existing user IDs. The data table consisted of rows with three values: the user ID that sent the data, the type of data, and the string containing the data that was sent. To access any piece of data we would then only have to query the data table for the rows with the correct user ID and type.

Tables	**Table Columns**
Data	User ID
	Type of Data
	The data as JSON object
ID	User ID
	Timestamp

Table 2 - Structure of database

4.2.2. The Data

Most data was stored as the string version of a JavaScript Object Notation (JSON) object, with some exceptions. A JSON object is a standard for formatting data, typically used when transferring data. A JSON object consists of attribute - value pairs, for example in the JSON objects we used for representing a piece of phone contact data we would have an attribute called "number" whose value is the phone number of that contact. The resulting JSON object would appear as follows:

{"Number": "123-456-7890"}

We stored the entire JSON object obtained from participants' phones so as to have as much information available as possible when we got to the machine learning stage. The following is how we stored each type of data:

- **Calendar, Text messages, Storage objects, and Call logs** - each of these data types were stored as the full JSON object that was obtained from the phone.
- **Contacts** - Each contact was stored as a JSON object with the contact's name and each of the phone numbers stored under the contact
- **PHQ-9** - The PHQ-9 was stored as a JSON object with keys one through nine corresponding to each question on the PHQ-9. The value stored under each key was the option (0 through 3) that was selected for the corresponding question.
- **Voice Recording** - recordings were saved as a .3gp file by Android. From there we converted it to a base64 string in order to send it over HTTP and store it in the database without errors.
- **GPS data** - GPS data was downloaded as a .kml file that simply listed the GPS location and times in tags. Therefore we simply saved the .kml file as a string.
- **Twitter Data** - We received Tweets from the Twitter API in the form of large JSON objects. The objects contained a multitude of information, including the tweet itself but also information about the poster's account, the GPS location of the poster when they posted, and how many likes and mentions the Tweet received. Therefore, we stored the entire Tweet as a JSON object in case any of the extra information correlated with depression

- **Instagram Data** - Similar to Twitter data, the Instagram data was also received as a large JSON object from the Instagram API. Instagram data was also stored as the full JSON object for the same reason as Twitter data.

4.2.3. Hosting the Server

The server was launched on a Worcester Polytechnic Institute (WPI) research virtual machine. The virtual machine required credentials to access, which only the students and faculty that participated in the study had. The database was a file on the virtual machine that could only be accessed from inside the virtual machine, so the data remained secure and only the project participants had access to during the study.

4.3. Machine Learning

The last piece of software that was developed was the machine learning testing environment. This software was implemented in Python version 3.5.3. There are four main parts to our machine learning architecture. These are restructuring the database, extracting features from the data, preprocessing the data, and training classifiers.

4.3.1. Restructuring the Database

The original database is a simple SQLite database that uses a simple database file to store data. This database often contains multiple entries for a certain person and modality. A script was written to convert this database into a folder of bytestream objects using Python's Pickle library, to create a single entry for a certain person and modality.

4.3.2. Extracting Features from the Data

For this step, some of the core libraries used were Numpy and JSON. Numpy functions were used in doing numerical computations, while the JSON library was used to access data that was stored in the JSON format. Examples for this type of data are Instagram and twitter data.

For conducting sentiment analysis, the TextBlob (0.15.1) library was used. For audio, we wrote a script that uses openSMILE, a free feature extraction tool that specializes on extracting

features from sound. To count faces in the Instagram modality, we needed to recognize faces. This was accomplished using OpenCV, which uses the Haar cascades technique.

4.3.3. Preprocessing

Preprocessing was done using scikit-learn's Imputer class, in addition to pandas and numpy to operate and manipulate matrices.

4.3.4. Training Classifiers

The classifiers used were all scikit-learn classifiers. Feature selection was done using the Randomized Lasso. Randomized lasso works by resampling the data and computing a Lasso on each resample. It is also known as the stability selection technique.

Hyperparameter optimization was done using GridSearchCV which stands for grid search cross validation. Grid search is a term for an algorithm that tries all possible combinations of provided hyperparameters for any learner, and reports back to you the best performing combination of hyperparameters. The cross validation comes into play when the success function of the grid search is calculated by using cross validation on the dataset.

For both of these concepts the scikit-learn implementation was used.

5. Experiments

In order to develop our application, we conducted a series of experiments in order to both better inform us about the problem we were attempting to solve, and to obtain data representative of the real world situation the application would encounter.

5.1. Exploratory Willingness to Share Study

As discussed in section 3.1, the purpose of the exploratory study was to determine how willing the general public is to give medical staff certain types of information. The survey was created using WPI Qualtrics, which is a tool for constructing surveys. We chose WPI Qualtrics because it is easy to quickly set up an attractive looking survey with logic for showing and hiding questions based on answers to previous questions. The Qualtrics survey was then conducted through Mechanical Turk in order to access the general public. We included a $0.05 compensation for every participant in order to encourage more people to take the survey.

The full survey can be seen in Appendix B, which can be used to aid in the following discussion of the survey. The survey began with an explanation of the survey. The explanation discussed the problem we were trying to solve: the need for a better depression detection method. The explanation discussed how responding to the exploratory study aids us develop an application that could improve depression screening. We included this explanation in order to encourage participants to give honest information, and to put the questions we ask in the study in context. The explanation next explained that the study is anonymous and does not actually gather any of the information types discussed, only their willingness to give the information. Participants are told that there are absolutely no risks to taking the survey.

After the explanation, the participants were confronted with the questions that make up the survey. The very first question was “Are you 18 or above?” If the participant was not older than 18, the survey immediately finishes without asking any other questions because using younger participants would have required extra IRB approval. If the participant was 18 or older, the survey continued on to a series of demographic questions. These questions asked participants’ gender, age, employment. The answers to the demographic questions were used to see if a certain group would be more or less willing than other groups to give information. Participants were also asked if they had ever been treated in an emergency clinic to see if they

have experience with the environment under which they would hypothetically be giving the information discussed in the exploratory survey.

After the demographic questions, participants were then asked about their willingness to provide certain types of information. For each of the remaining questions asked in this section, participants were asked to label their willingness as “completely unwilling”, “somewhat unwilling”, “unsure”, “somewhat willing”, or “completely willing.” We felt that the five options would allow us to gauge what scale of willingness the participant felt, not just willing or unwilling but to what degree. The first group of questions in the section contained questions about giving information from social media accounts. The four questions then asked the participants willingness to provide medical staff with their Twitter username, their tweets on Twitter, their posts on Facebook, and their messages on Apps such as GroupeMe, Discord, and WhatsApp. The second group of questions considered giving data stored on a typical smartphone. Participants were asked about their willingness to give medical staff their phone’s GPS data, gyroscope and accelerometer data, browser history, call logs, and App usage data. The last group of questions in the section asked about willingness to perform certain tasks while being recorded. The two questions asked were how comfortable the participant was to speak a phrase into a microphone, and how comfortable they would feel allow an image of their face to be captured and analyzed. We felt that grouping the willingness questions into three smaller groups provided a more clear and organized survey that participants could follow easier.

After the willingness questions, participants were asked their age again, but this time with the options reversed. We included this question a second time in order to check the results of the survey for thoughtless results were the participant quickly marked answers without reading the questions, and for automatic response, or “bots.” Amazon Mechanical Turk has had problems with many of their survey takers being bots in the past, which do not give good data. By including the same question twice we could check the results to see if the participant chose different ages which would indicate a bad respondent.

Finally, we gave a blank text field were participants could leave feedback or comments. We included this section in order to gather extra opinions of participants if they had strong feelings on the subject which could also be considered.

5.2. Data Gathering Study

The goal of the data gathering study was to obtain a large set of data with corresponding PHQ-9 scores that could be used to train a machine learning model. The data gathering study was a study conducted on Amazon Mechanical Turk in order to gather a base of data from the general public's phone to user to train machine learning systems. In order to conduct this study, two pieces of software were developed: an Android application and a server for receiving and storing data.

5.2.1. Incentives and Design Choices

As previously mentioned, the study was conducted on the Amazon Mechanical Turk platform. The use of this platform provided several benefits to our data collection process. Firstly, it allowed us to get cost effective data from hundreds of participants. At a base cost of $0.40 per participant, we correctly anticipated hundreds of responses before we would deplete our budget. This was an important factor, as our machine learning algorithms required a very sizeable dataset to be properly trained. In order to encourage users to give more personal data, we incentivized users for providing Audio and online account data (Google, Twitter, and Instagram). The users would be provided a "bonus" for providing these data types, which constituted $0.10 each for Audio, Twitter and Instagram data, and $0.30 for Google data. The Google data was strongly encouraged through the bonuses, since it contained highly useful GPS data.

Mechanical Turk also allowed us to set requirements for potential participants wanting to complete our study. Since we wanted a large dataset, we decided to set a minimal requirement of having successfully completed 50 other tasks on Mechanical Turk. This requirement meant that people taking our study would have experience with the platform, and that their work has been accepted by at least 50 other Requesters, meaning it has some proven value. This was anticipated to improve the quality of the responses we would receive.

5.2.2. MTurk Project Parameters

In setting up our project on Mechanical Turk, there were several parameters that needed to be defined. For the time allotted per assignment, we set 2 hours to give participants enough time to install and complete our app. We also set 600 as the number of unique participants in our first batch. However, we quickly realized that as the survey progressed and our study disappeared off the first few pages, we saw a significant decline in the number of participants.

5.2.3. Result Filtering

In order to confirm accuracy of the results gathered from our participants, the results were tested with several validation scripts. First, we needed to cross check the validation codes provided by the app with the ones in the database. If a code didn't match with any existing record in the database, the participant was contacted and asked to provide a valid code. If a valid code was provided, the script proceeded to check if the database does indeed contain all the data types that the completion code specified. If an error was found at this point, the participant was contacted and asked to complete the application again with all the correct data types. Finally, the script would then automatically assign the appropriate worker bonuses to each of the workers that provided data types that guaranteed additional compensation, such as GPS, Twitter, Instagram and voice data. A separate script would then take all the validated ID's and run them through an algorithm that would generate on-the-fly statistics about the current state of the data.

This setup allowed for quick and easy analysis of the study data as it came in. This allowed us to immediately start creating machine learning systems based on the preliminary results, before the survey was complete.

5.3. Machine Learning

The machine learning experiments consisted of seven distinct parts. These were conducting regression tasks, converting the PHQ-9 label into a binary label for depressed and non-depressed for certain cutoffs, and training learners that use only a certain modality for its dataset, dataset balancing, feature reduction, and training different kinds of learners. After all these parts, the method of bagging and consequently a meta-cost learning approach was also explored to possibly produce learners that generalize and fit our use case better.

5.3.1. Regression of PHQ-9 Score

Our dataset, in its rawest form, presents a depression label as a PHQ-9 score with associated user data. This allowed for input into regression learners, which predict continuous values.

5.3.2. Binary Classification of Depression

The PHQ-9 survey consists of 9 questions, each of which can contribute to a final score of depression with 3 points, resulting in a depression score that ranges from 0 to 27. It must be noted that the official cutoff for depression set by Pfizer, the creator of the PHQ-9 survey is 10. There are other cutoffs, like 10-15 for mild to moderate depression, 15-20 for moderate to severe depression, and 20-27 for severe depression.

The robustness of these cutoffs have been put to test by Kroenke et. al (Kroenke et. al, 2001), and it has been found that a PHQ-9 score bigger than 10 has a sensitivity of 88% for major depression.

In our project, we tested our best performing classifiers for cutoffs of 10, 15 and 20, trying to see if we can train learners that can robustly detect certain levels of depression.

5.3.3. Modality Based Correlation

Our project is unique by the fact that it tries to use multiple modalities to predict depression. Since our use case allows for a user to provide a subset of the modalities that has been the subject of our discussion, we felt it was wise to work on learners that learned the relationships between our depression label and a certain modality.

All the different types of labels, be it the continuous type or the binary type at a certain cutoff, were utilized as labels for each and every modality. These individual learners were then combined to form an ensemble classifier that could work with any subset of modalities.

5.3.4. Dataset Balancing

The data that we've gathered is sometimes particularly skewed at certain PHQ-9 cutoffs. For example, the 20 cutoff yielded around 30 people with severe depression, and the remaining 300 have non-severe depression or no depression at all.

In cases like this, our dataset skew introduced bias to the learning task at hand. In addressing this issue, we found it wise to balance the dataset ie. make it so that in binary labeling, the number of positive and negative labels matched each other in numbers.

Although this reduced the number of samples that were available to us in certain configurations of the dataset, ultimately it allowed us to achieve precision and recall scores that showed no sign of bias.

5.3.5. Feature Reduction

It is common knowledge in the subject of learning that oftentimes, too many features insignificant predictive value introduces noise and make it so that it is harder for the classifier or the regressor to find a set of solutions that achieves a good local minima.

It is also known that feature reduction allows for learners to generalize over the target population with more success ie. reduce overfitting.

The best way to combat this is to reduce the number of features by using some type of feature selection method or an importance metric ie. assigning importance to each and every feature, and selecting only the top features, or the features that have non-zero importance weights.

To carry out this tasks, variance tests, chi squared tests, using the inbuilt feature importance attribute of a random forest learner, or using stability selection with l1 regularization can be utilized.

For our project, our best performing classifiers were often Support Vector Machines that used radial or linear kernels. Kernels in SVM are different ways to calculate the distance between the support vector and a datapoint. Different kernels convert the distance calculation into different spaces.

We found it sensible to use stability selection with l1 regularization or a random forest classifier, since both of these methods provide some type of selectiveness on the dataset by enforcing a metric of information gain.

5.3.6. Comparing Various Machine Learning Algorithm

Our experiments early on with modality based classification, that is bijecting our dataset into modalities and experimenting on specific modalities, and regression includes linear

regression, random forest classifier, random forest regressor, SVM for classification, SVM for regression, a basic multi-layer perceptron with one hidden layer (neural network), logistic regression, naive bayes and knn.

As we observed the successes of these different learning algorithms, only SVM for classification, SVM for regression and random forest classifier and regressor proved to be useful in our quest to find the best learner for each modality.

5.3.7. Bagging and Meta-Cost Learning

The aim of machine learning is to produce learners that generalize their intended task over unknown data. For a classifier to be able to do this better, bagging is a method that can be utilized. In bagging, many instances of a particular learner are run on random subsets of the datasets, decreasing variability and possibly reducing over fitting as a result.

We also use meta-cost learning to make our final learners fit our use case of detecting depression better. In our use case, the penalty for not detecting depression can possibly result in fatality, whereas in the penalty of misdiagnosing someone not depressed as depressed will only put mild to moderate financial stress on the participant. With this in mind, we optimized our final learners for increased recall for the depressed label.

6. Results

6.1. Exploratory Willingness to Share Study

As discussed in the sections 3.1 and 5.19, the exploratory study asked participants about their willingness to provide medical staff with certain types of information. From the exploratory study, we were able to deduce what data a medical professional would likely obtain from a member of the general public.

6.1.1. Social Media

When it came to Twitter, we found that 44.64% of participants were willing to provide their Twitter username and 44.2% were unwilling, as seen in Table 3. Table 4 shows that we found that 50.45% of participants were willing to allow a member of the medical staff to retrieve their tweets, compared to the 40.18% that were unwilling to some degree. The results suggest that the population is split in approximately two halves when it comes to sharing Twitter information. It is important to note that as stated in the Background section, about 24% of the population actually has a Twitter account, so some participants might not have an account but simply assume they would not want to share the account if they made one. Therefore, when taking this into considerations and comparing the results for Twitter to the rest if our results we believe that participants were fairly willing to provide their Twitter data and would be feasible to obtain.

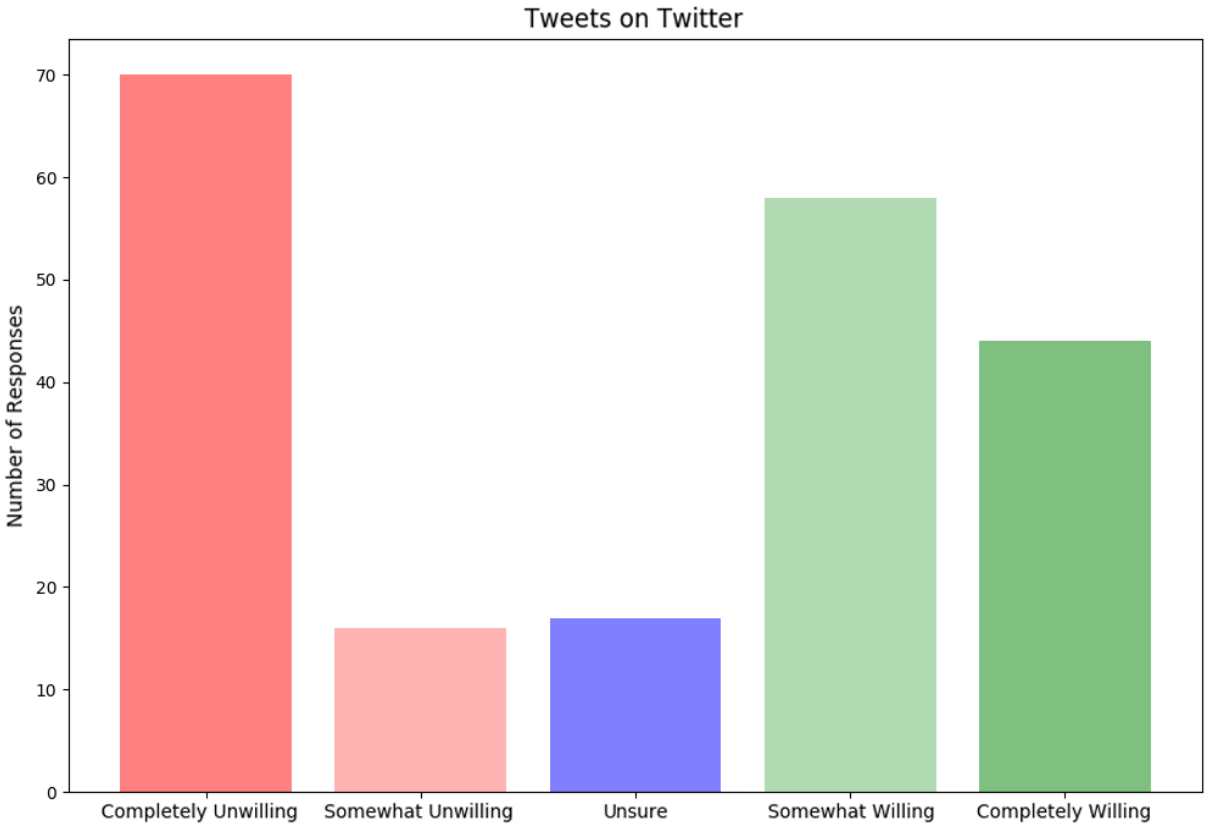

Table 3 - Participant willingness to share their Twitter username with a medical professional

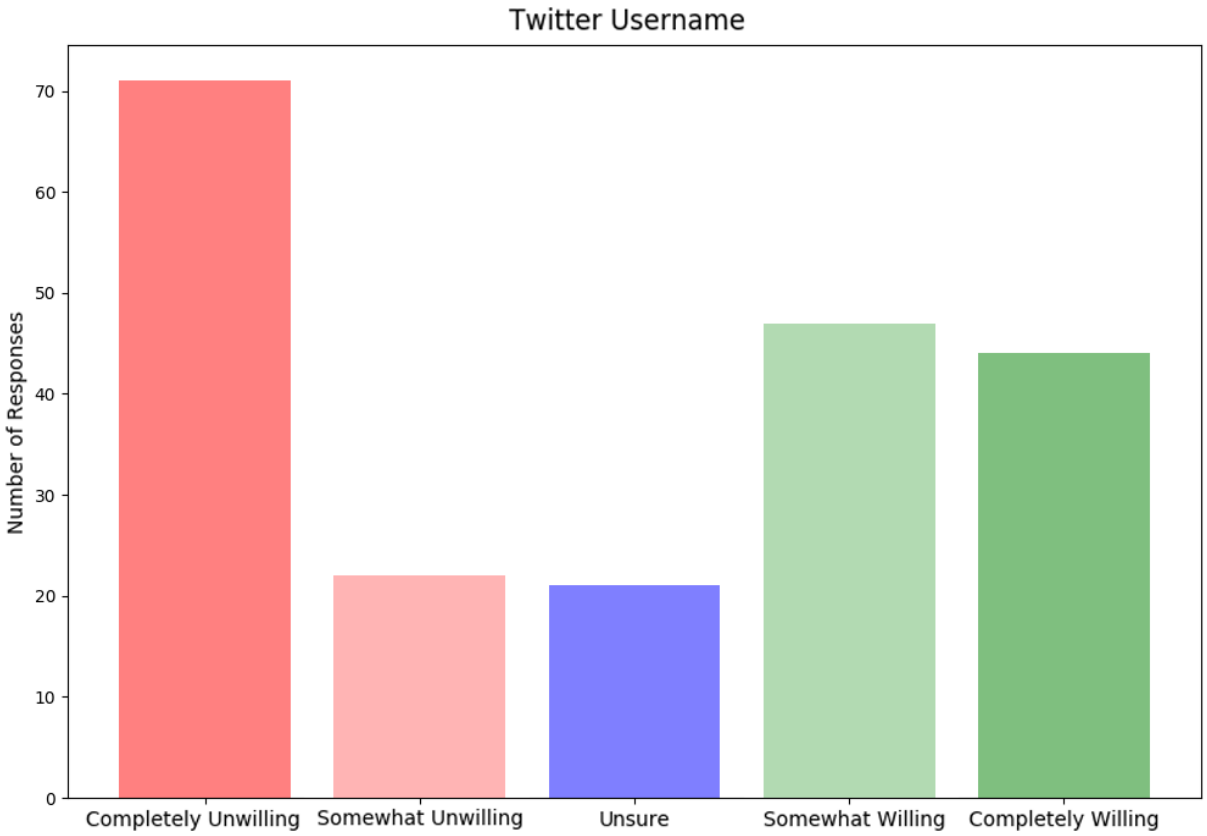

Table 4 - Participant willingness to share their tweets on Twitter with a medical professional

We found that 45.09% of participants would allow a medical professional to access their Facebook posts compared to the 42.41% who would not, which is shown in Table 5. However, by this point we also discovered that it would be prohibitively difficult to make an automated system for access Facebook posts. We would have had to store users' passwords and usernames which we felt would be a breach of privacy. As shown in Table 6, we found that for platforms such as Groupme, Discord, WhatsApp, 43.75% of participants chose were willing to some degree to provide the data compared to the 47.32% who were unwilling. Again, the results show the population is generally split, but these results are slightly lower than the results of other questions.

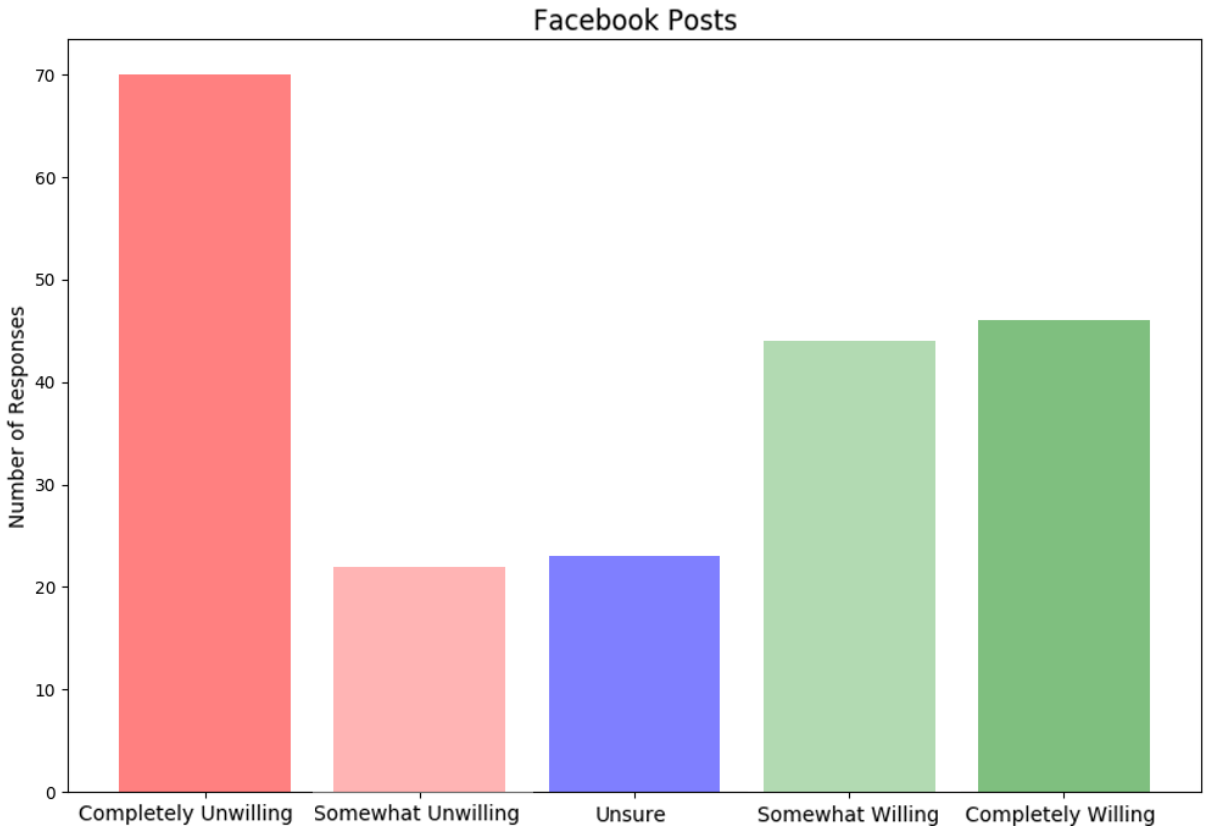

Table 5 - Participant willingness to share their Facebook posts with a medical professional

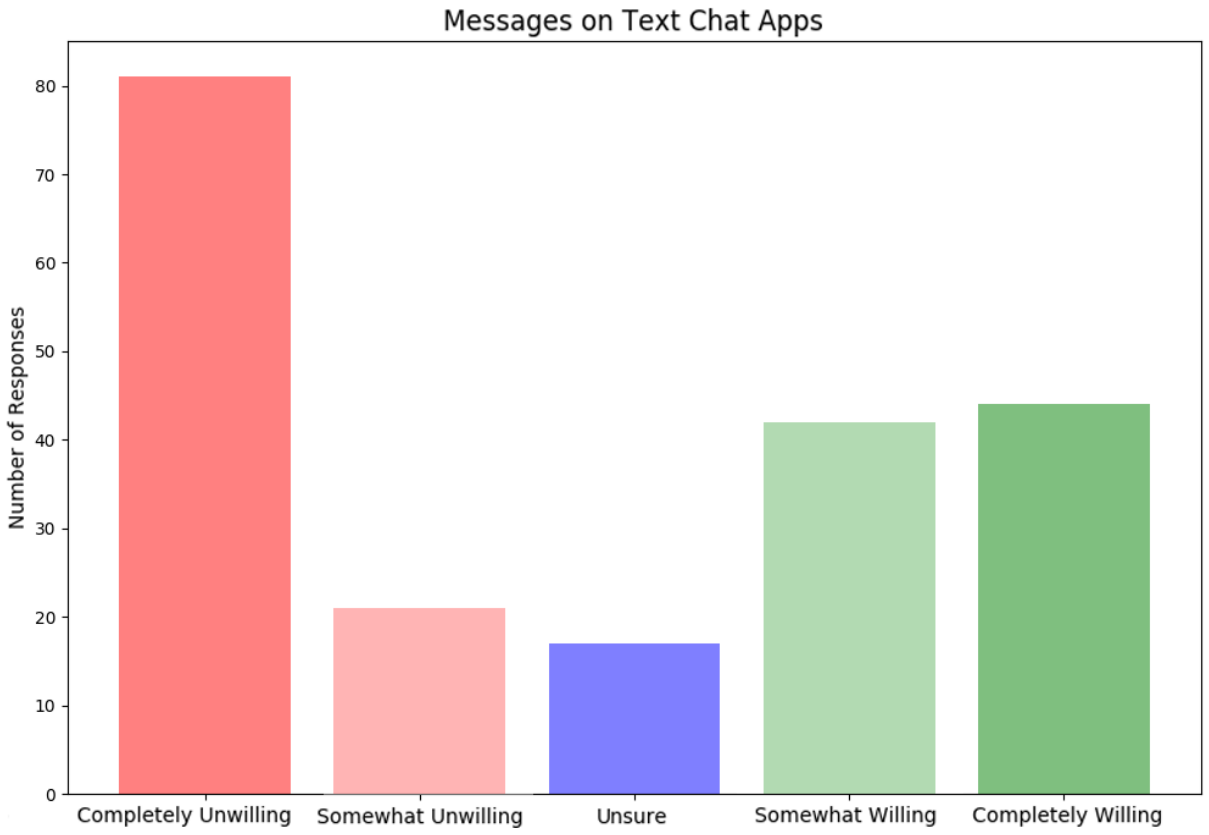

Table 6 - Participant willingness to share their messages on text chat apps such as GroupMe, Discord, or WhatsApp with a medical professional

6.1.2. Phone Data

When asked if they were willing to provide medical professionals with their historic GPS data, 47.80% of participants replied that yes they would be some degree of willing as opposed to the 41.95% who were not willing, as seen in Table 7. We were surprised to find to see that not only did a majority of participants feel comfortable sharing their location, but by a reasonably strong margin when compared to our other results. Participants in the study were less willing to give gyroscope and accelerometer data, with only 44.39% willing to compared to the 43.41% that were not willing to. As shown in Table 8, the results were split evenly. However, when it came to browser history we found that participants were much less willing to share their data with medical professionals. Only 37.07% of participants were willing to share their browser history compared to the 53.17% who were not, as seen in Table 9. The results match our expected results, since browser history is often thought of as being very private. Regardless, we had already discovered that we were incapable of retrieving the browser history from Android

phones through our application. Participants were also rather unwilling to provide their call logs, as seen in Table 10. Only 42.93% of participants were willing, and 47.32% were unwilling. Finally, participants were more willing to provide their app usage data, which is how often and how long certain apps are open on their phone. Table 11 shows that 47.32% of participants were willing to provide medical professionals with their app usage data, which 44.39% were not willing to. We felt that while call log's results were not great they were still good enough to be feasible to obtain. However, we discovered the privileges required to obtain app usage data were too difficult for us to obtain, so we would not be able to use it.

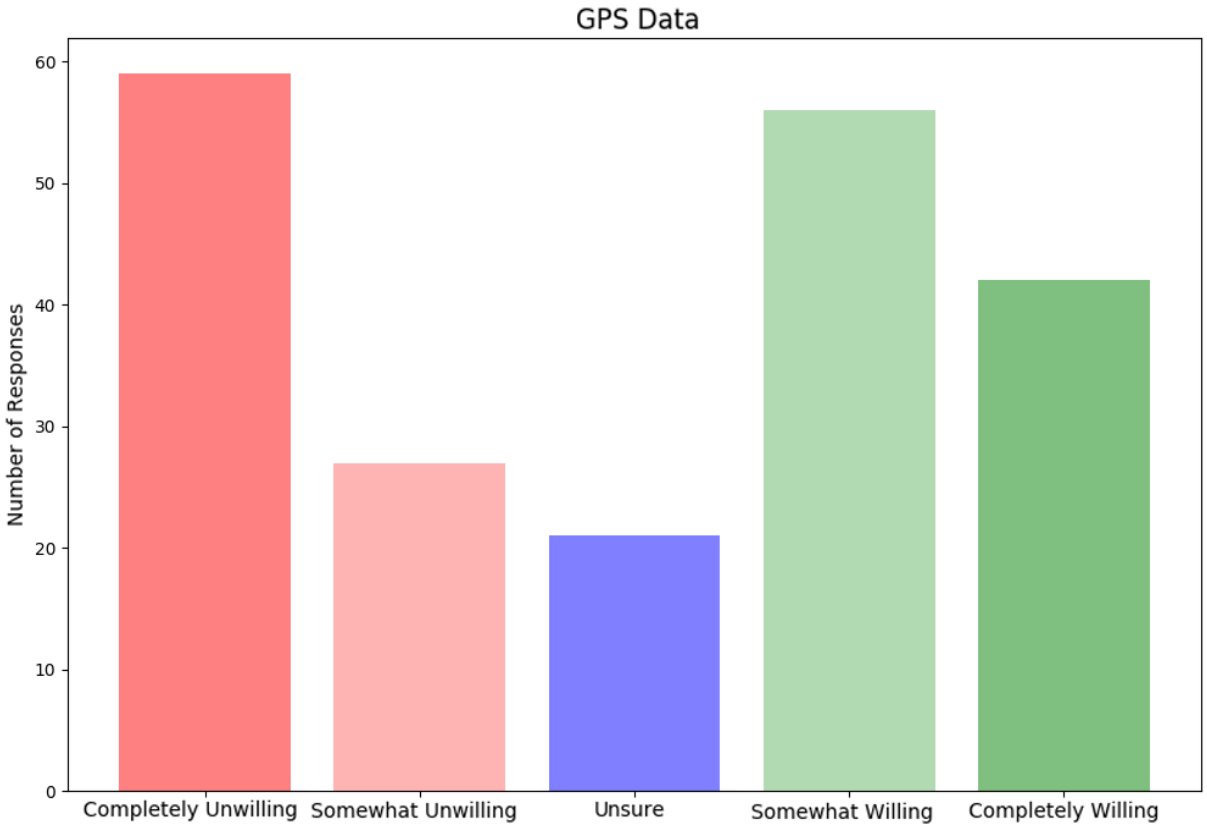

Table 7 - Participant willingness to share their historic GPS data of the last two weeks with a medical professional

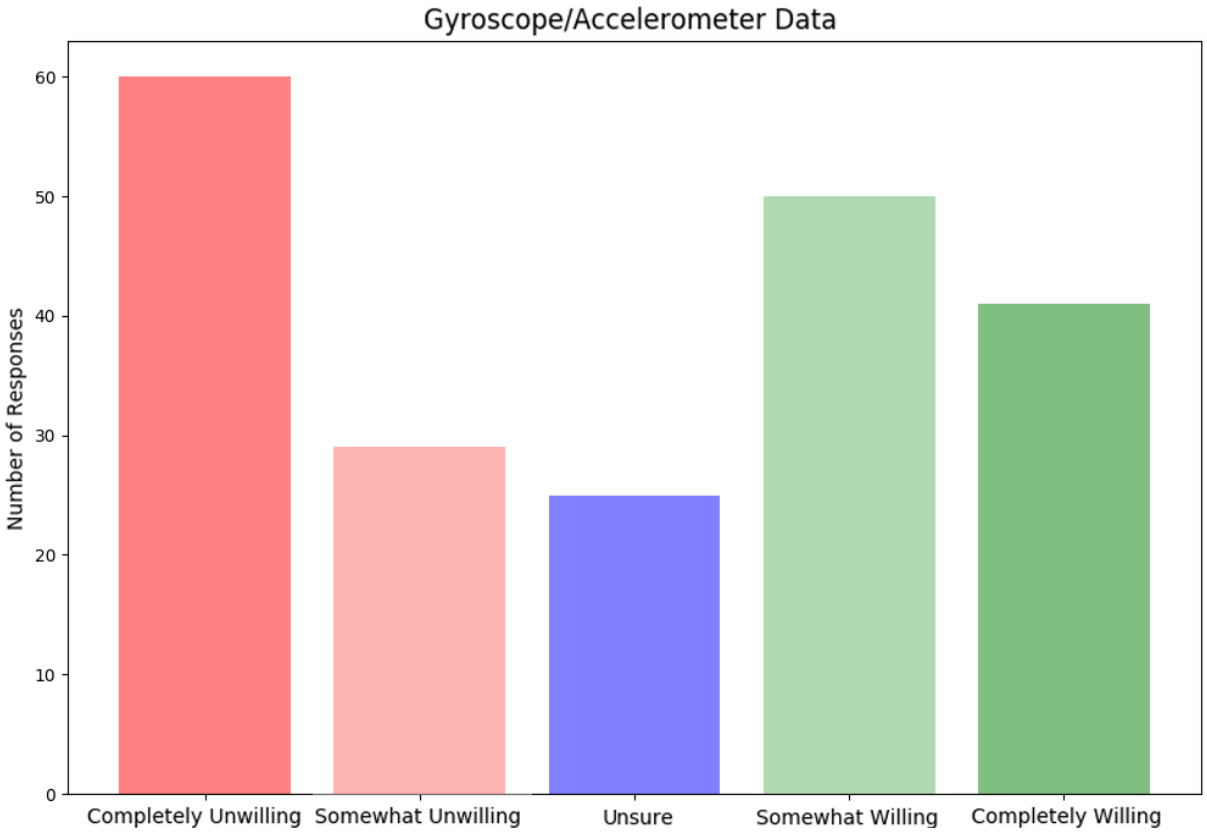

Table 8 - Participant willingness to share smartphone gyroscope and accelerometer sensor data with a medical professional

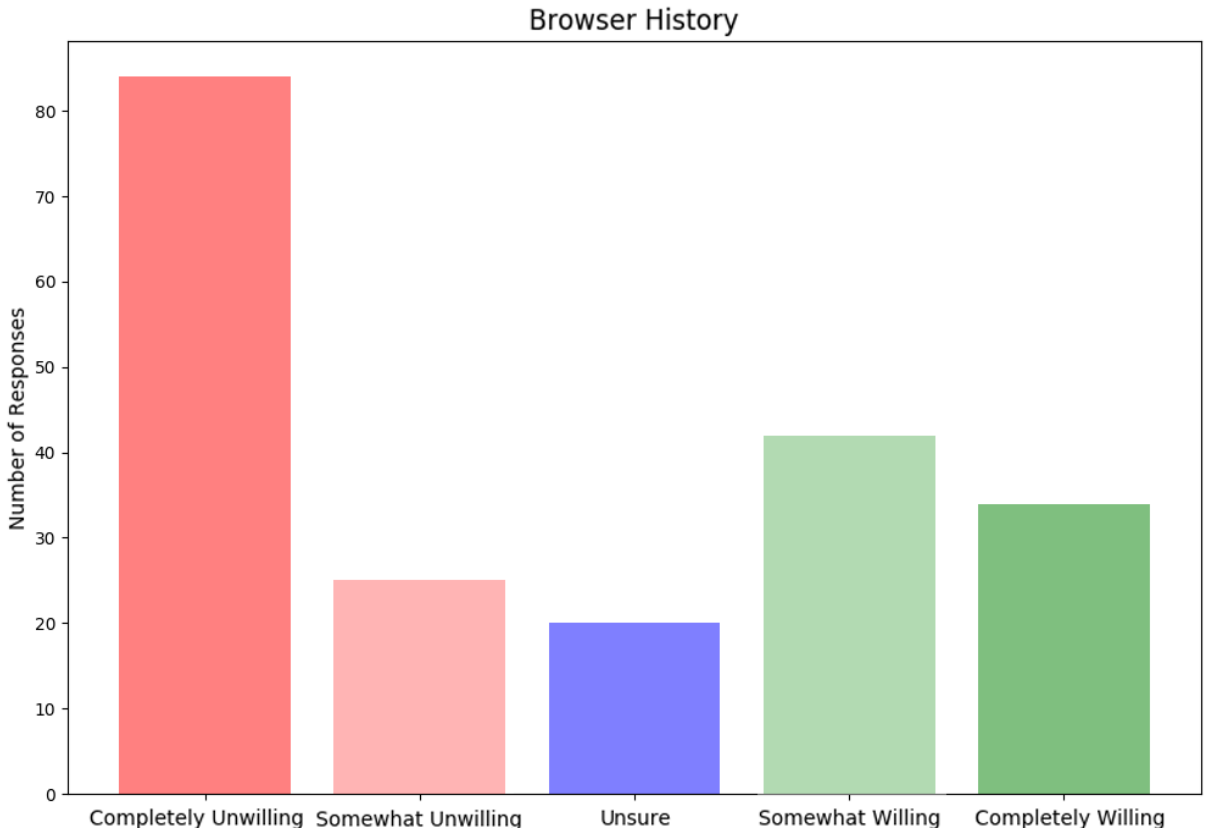

Table 9 - Participant willingness to share their browser history with a medical professional

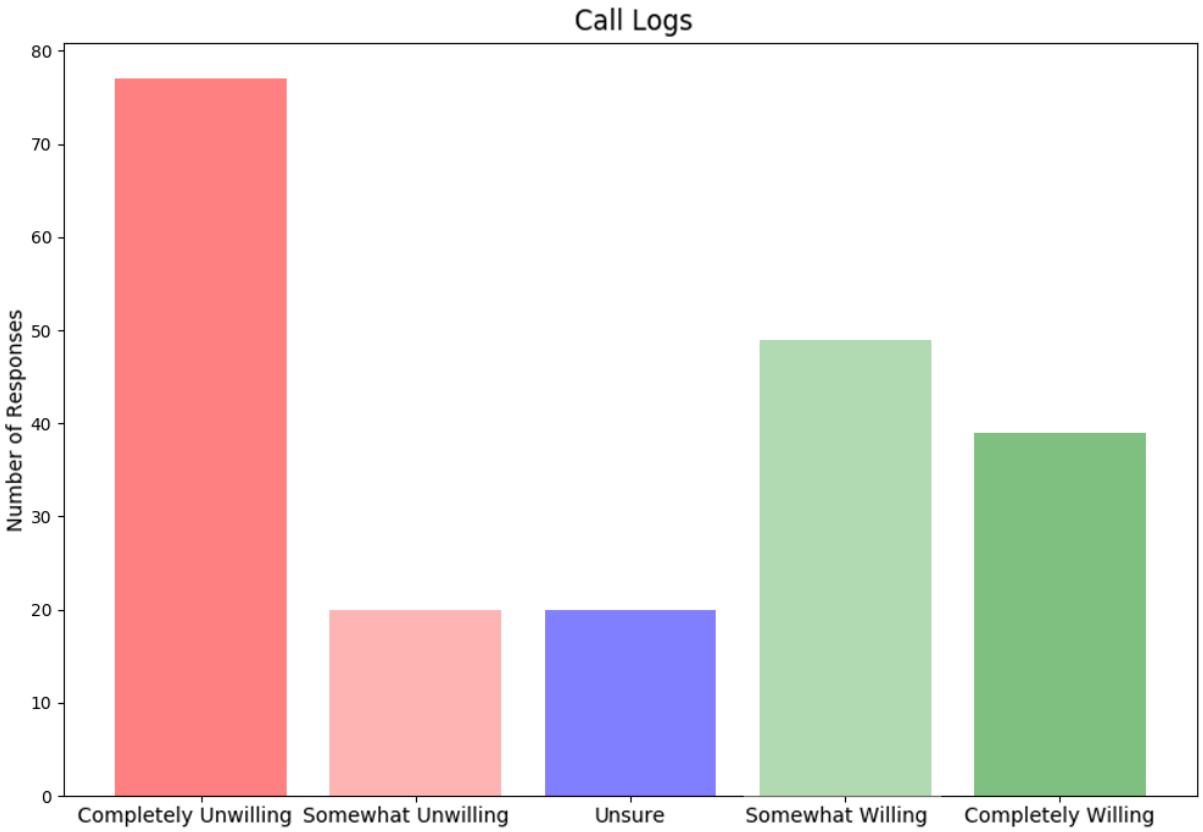

Table 10 - Participant willingness to share their call logs with a medical professional

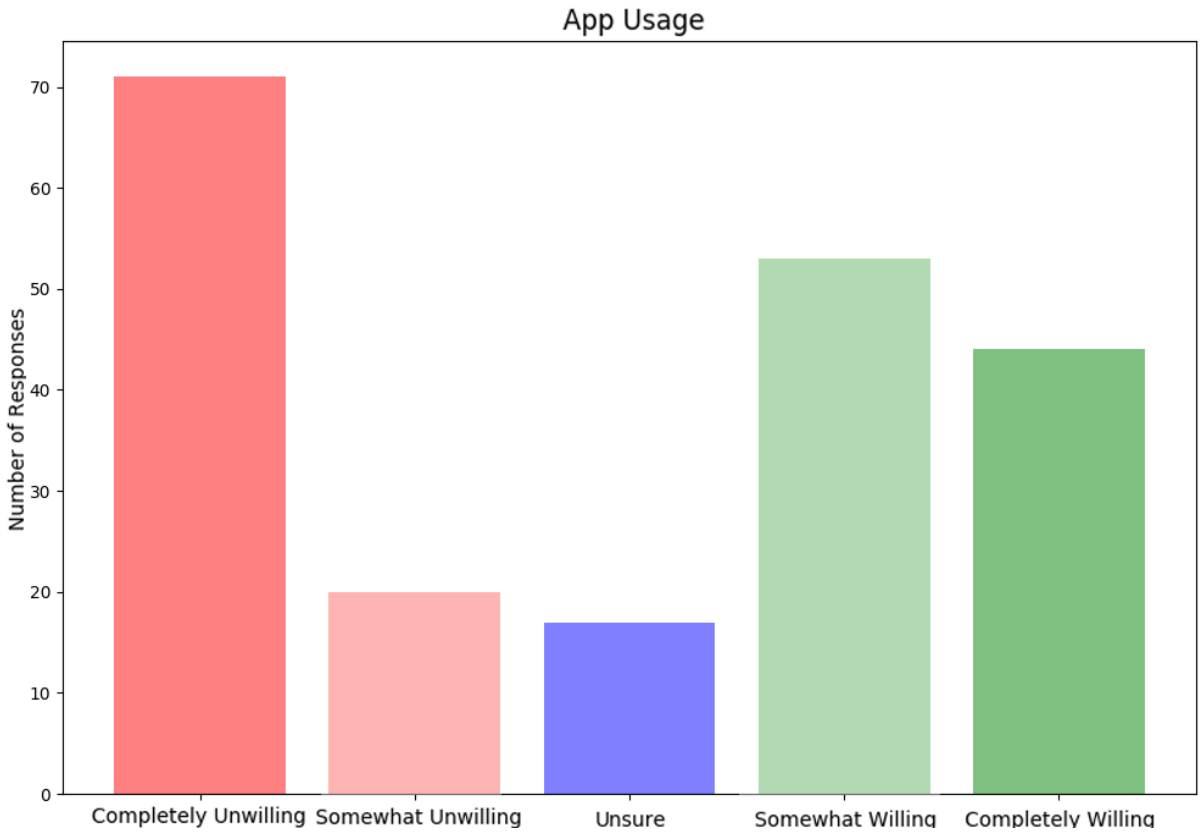

Table 11 - Participant willingness to share their app usage data with a medical professional

6.1.3. Recording Voice and Face

We found that 63.39% of the participants were either extremely or somewhat comfortable to speak a phrase into a microphone, compare to only 20.09% we either extremely or somewhat uncomfortable performing the task. The study also showed that about 57.59% were either extremely or somewhat comfortable permitting medical staff to capture images of their face, compared to 27.24% of the participants who were either extremely or somewhat uncomfortable. Both of these results are very positive, showing the highest willingness scores of any question asked. Therefore, we found that people are typically more willing to allow recordings of themselves in the moment than allow historic recordings of data stored on their phone and online accounts.

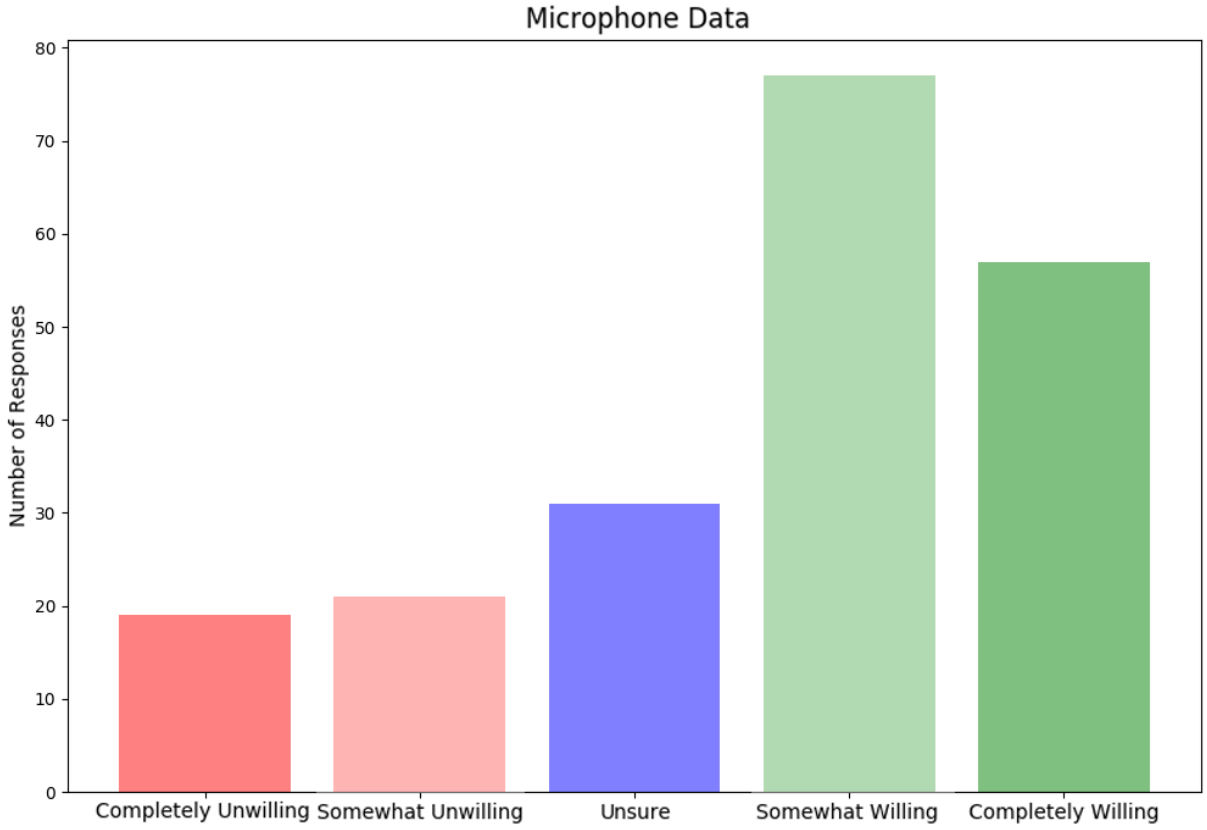

Table 12 - Participant willingness to share their microphone data with a medical professional

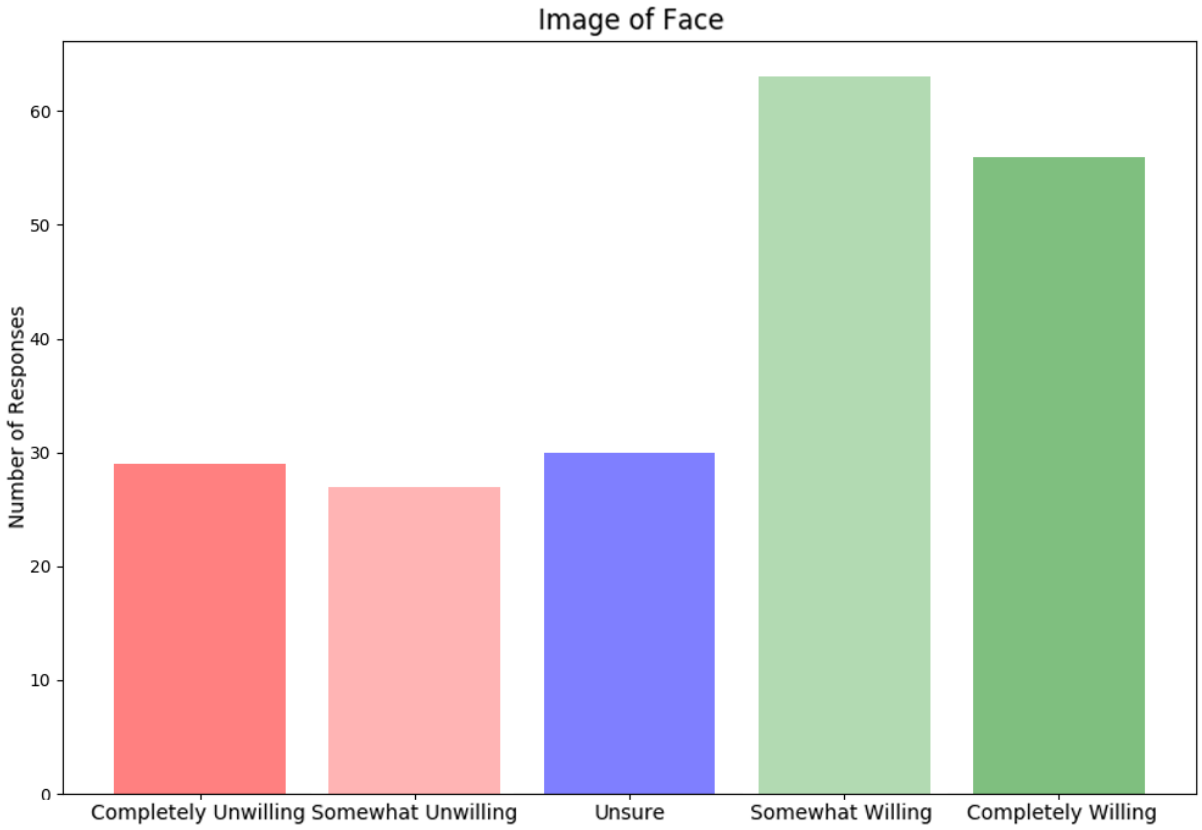

Table 13 - Participant willingness to share their facial images and data with a medical professional

The most useful observation made from the willingness study was that people are expectedly hesitant to provide very private information such as their text conversations and browser history. While you could consider a person's voice and face to be a very personal and identifiable thing to provide to an application, most people use their voice and face to interact with the world in day to day life, so they are more desensitized to people observing this data. Also, a picture of a face and a voice recording are more immediate and limited in scope, while a patient could have text messages stored on their phones that go back several years, and could provide a much more concrete and personal look into the person's private life, which might cause unease.

6.2. Data Gathering Study

The MTurk Survey completed with a total of 414 responses. Out of these responses, 73 were rejected from the final dataset due to incomplete or bogus data. Due to the filtering process,

the final dataset used with the machine learning models consisted of 341 entries. The final payout for the data gathering study consisted of $350.77, with an average of $0.85 per participant. Each participant was provided a basic compensation for providing easy-to-extract phone data, and was additionally compensated for each other modality. The following table presents a breakdown of our compensation per modality:

Modality	**Payout**
Basic phone data (texts, call logs, contacts..)	$0.40 **(Base payout)**
Twitter	$0.10
Google GPS Data	$0.30
Instagram	$0.10
Voice Recording	$0.10

Table 14 - Participant compensation per modality

Since our data validation scripts were continually improved upon over the course of the study, some participants were accepted and compensated, only to later be rejected due to discovery of fake or unusable data.

6.2.1. Data Type Breakdown

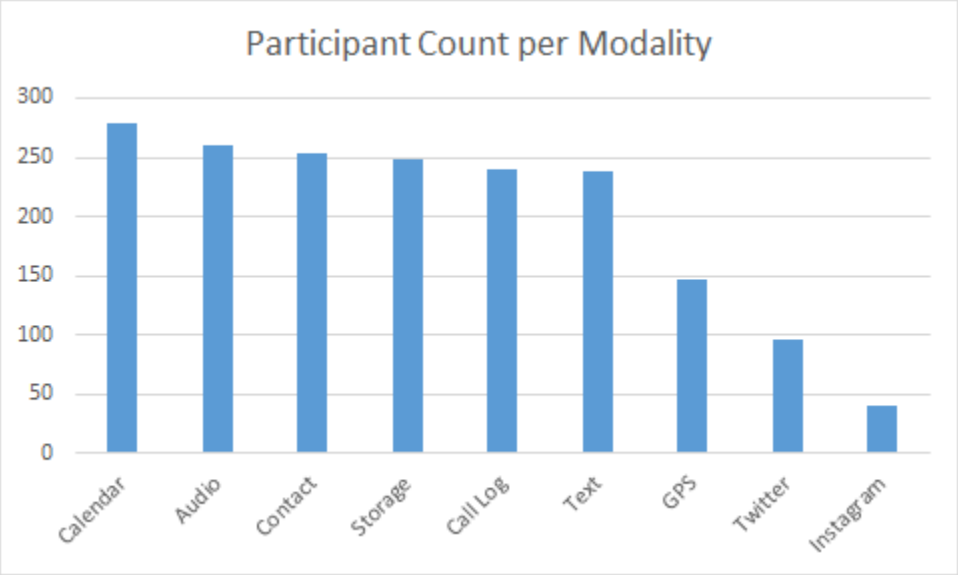

Table 15 - Number of participants per data type (modality)

As seen in Table 15, the number of data points along each modality ended up being very similar, with most categories being in the 239-278 range of responses. The data type that users found most comfortable sharing was Calendar data (81.5%), followed by Audio (76.5%), Contacts (74.2%), Storage (73.0%), Call Logs (70.4%) and Texts (70.1%). The three least provided data types were GPS (43.1%), Twitter (28.2%) and Instagram (12.0%).

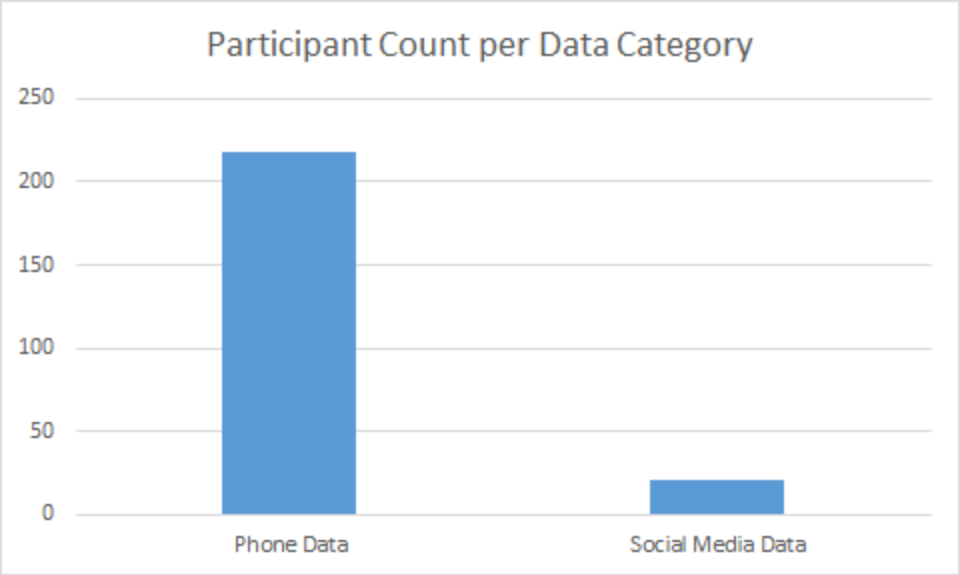

Table 16 - Number of participants per data category

When it comes to the five essential data types we collected (calendar, contacts, call logs, texts and storage), 218 (63.9%) of the participants were willing to provide all five of the different data types at once (see Table 14). Conversely, only 21 (6.2%) of participants provided all three online account related data types (Google GPS, Twitter and Instagram), though this is likely related to the low crossover between users of these three services in general.

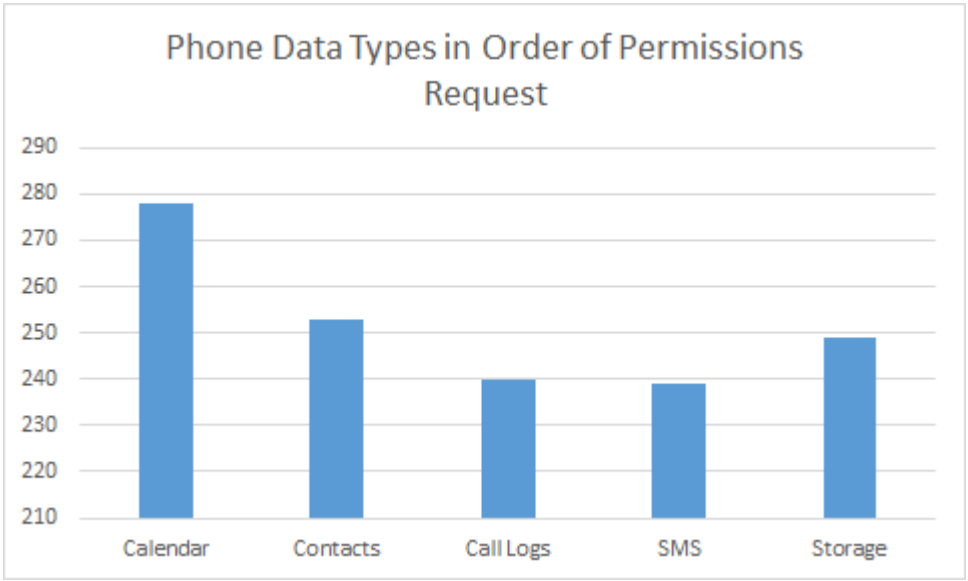

Table 17 - Phone data types sorted by the order in which they are requested for permission to gather from the user

The number of responses shows a decrease as more and more permissions are requested from the user to provide access to their private data, as shown in Table 17. 278 of the participants were willing to very willing to provide the calendar data, which was also the first requested. However, as the additional permissions were requested, a downward trend appeared, dropping to 239 when the text request appeared, and then edging slightly higher on the storage request to 249.

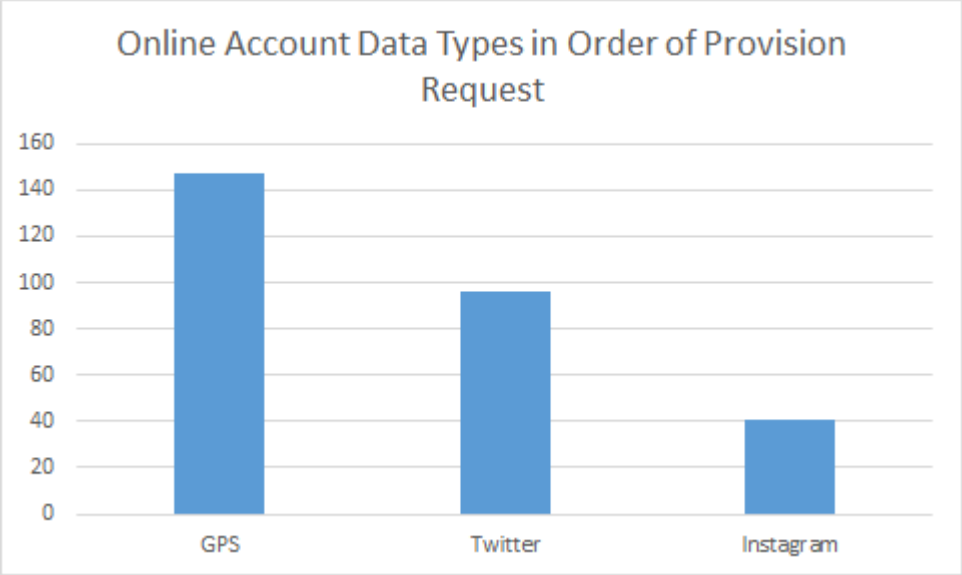

Table 18 - Online account data types sorted by the order in which they are requested for permission to gather from the user

Similarly, Table 18 shows that the online account data types had the highest number of responses (147) for the first requested, which was Google, and the least number of responses (41) for the last requested, which was Instagram.

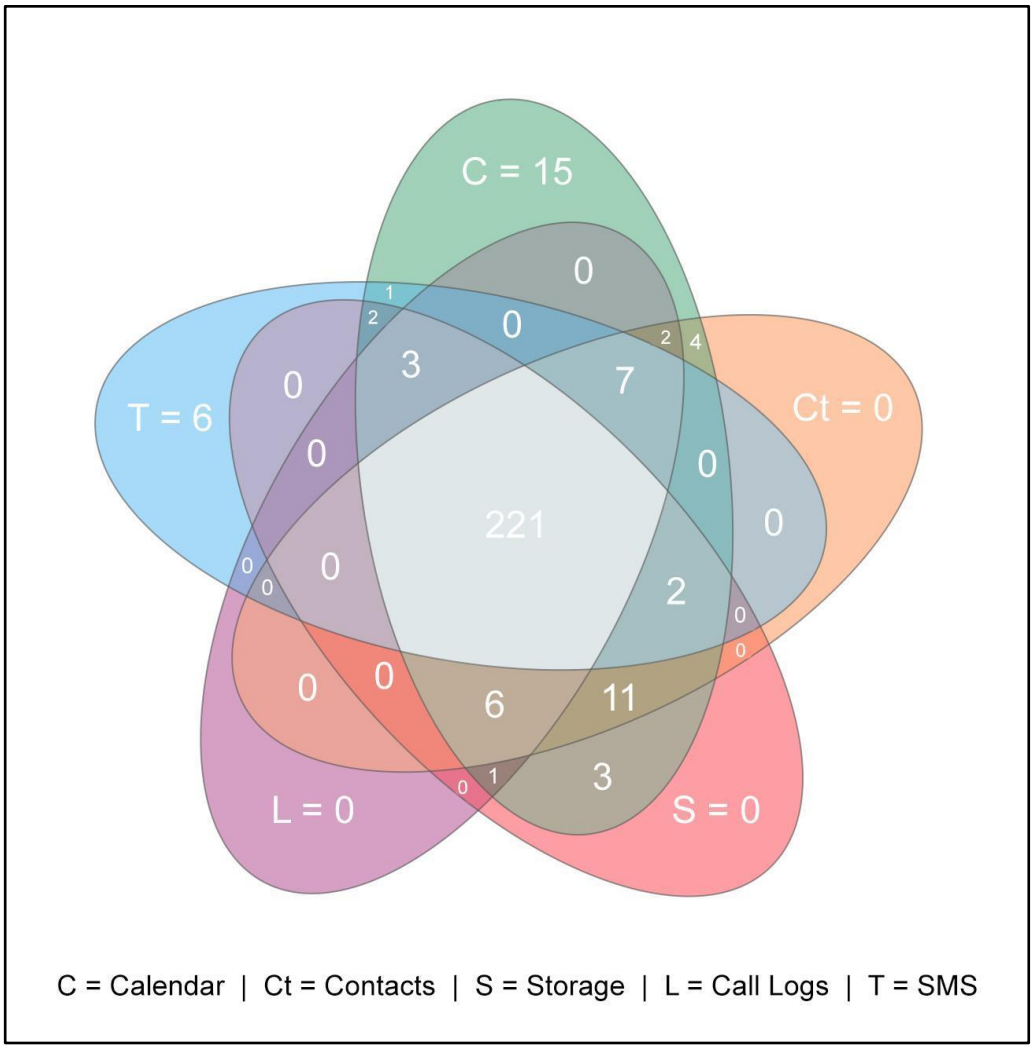

Figure 21 - Breakdown of responses by data type

As shown by Figure 21, most users were willing to provide all five types of basic data. It also shows the skew towards the first asked permission (calendar). Out of those who provided any basic phone data, only 6 out of 284 were unwilling to provide access to their calendar data. Interestingly, those 6 participants provided access only to their text message data.

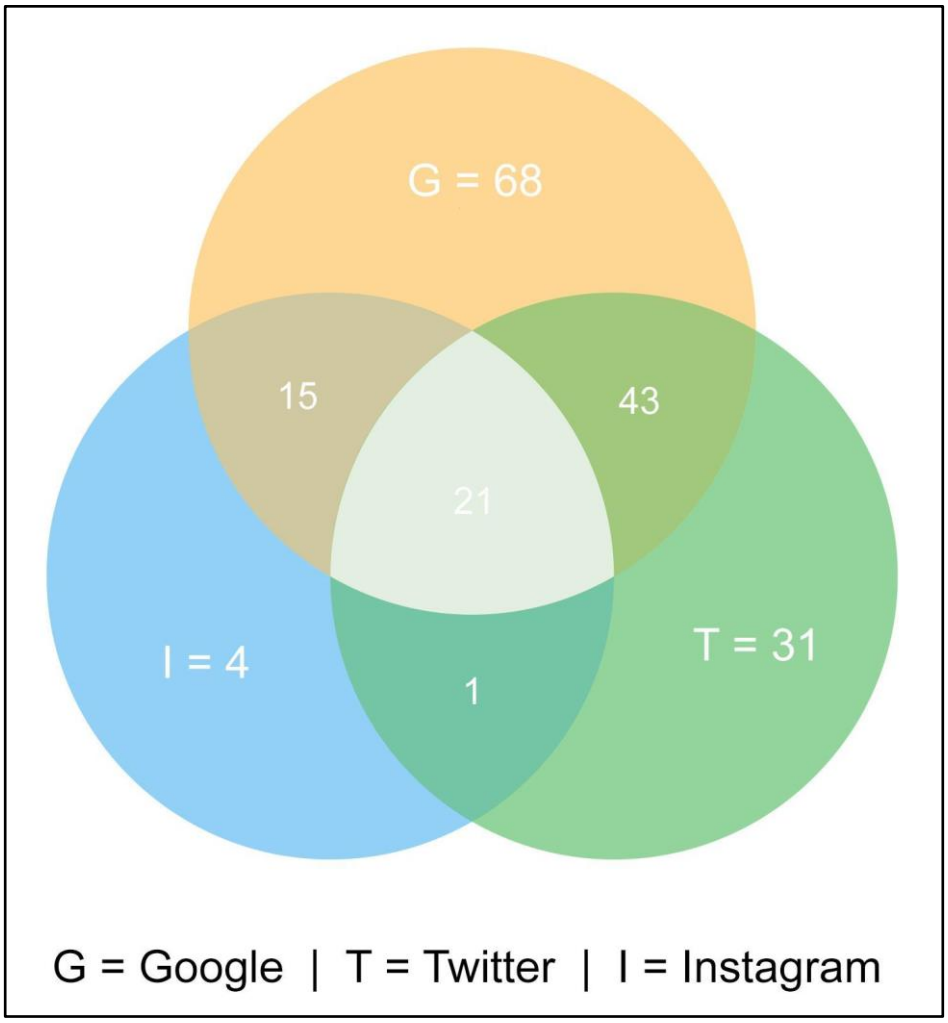

Figure 22 - Breakdown of responses by online account data type

As previously mentioned, Figure 22 shows that Google data was the most likely to be provided from the online account data. Only 11.5% of users that submitted some type of online account data submitted all three different types.

6.3. Machine Learning

As mentioned earlier in the experiments section, our experimentation with machine learning methods consisted of six distinct parts. These were conducting regression tasks, converting the PHQ-9 label into a binary label for depressed and non-depressed for certain cutoffs, and training learners that use only a certain modality for its dataset, dataset balancing, feature reduction, and training different kinds of learners.

Through conducting these six steps at various stages of development, we have arrived at various results. They are presented below.

6.3.1. Regression of PHQ-9 Score

Our labels were aggregate PHQ-9 scores in their raw form, thus we experimented with regressors first. We tried out many regressors, and for all modalities, the support vector machine regressor continuously provided robust results. For all modalities, hyperparameter optimization was conducted, and mean squared errors on both the training and testing set for these optimized classifiers are, upon undergoing a square root operation thus as mean errors, provided below:

	Audio	Instagram	Text	Contacts	Twitter	Call	GPS
Training Set (4-FoldCross Validation)	5.39	6.32	6.78	7.14	7.07	7	6.55
Testing Set (2-Fold Cross Validation)	7	1.41*	6.24	6.56	5.74	6.16	6.63

Table 19: Regression results (RMSE) for all modalities

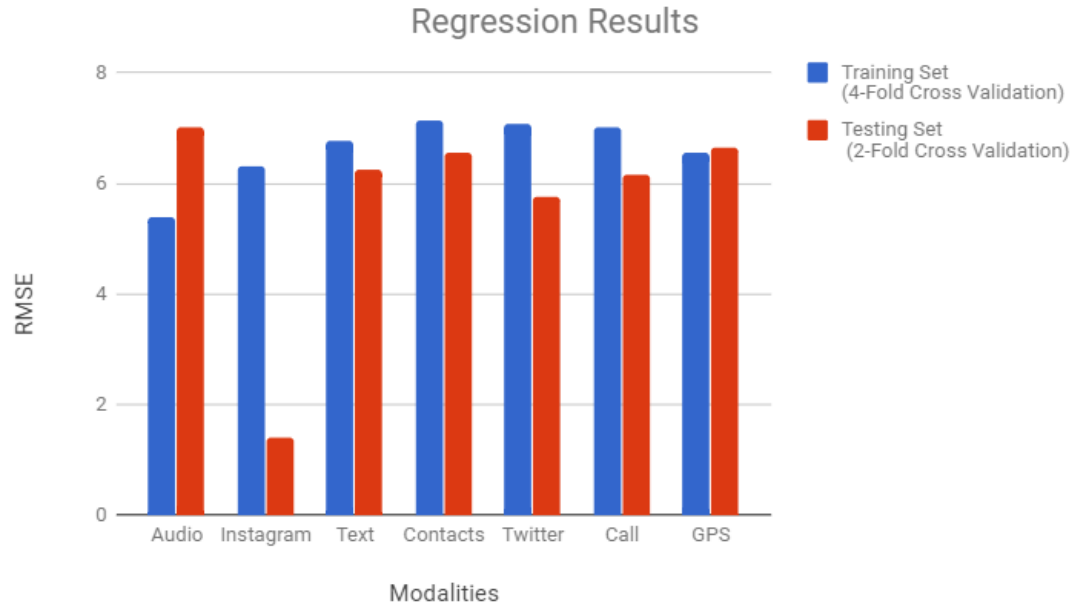

Figure 23: Regression results

It must be noted that the Instagram testing set error is extremely low. This can be attributed to the fact that since the Instagram modality response rate was very low, and our testing set encompassed 15% of the entirety of the data, the Instagram test set had around 4 data points.

It should also be noted that a mean error of 7, which is the worst mean error we achieved, is roughly equivalent to 70% accuracy.

6.3.2. Binary Classification of Depression with Different Learners

For binary classification of depression, we considered multiple PHQ-9 score cutoffs for creating 2 bins (depressed vs not-depressed). Table 17 includes accuracies of SVM classifiers ran on the combination of certain PHQ-9 score cutoffs and modalities.

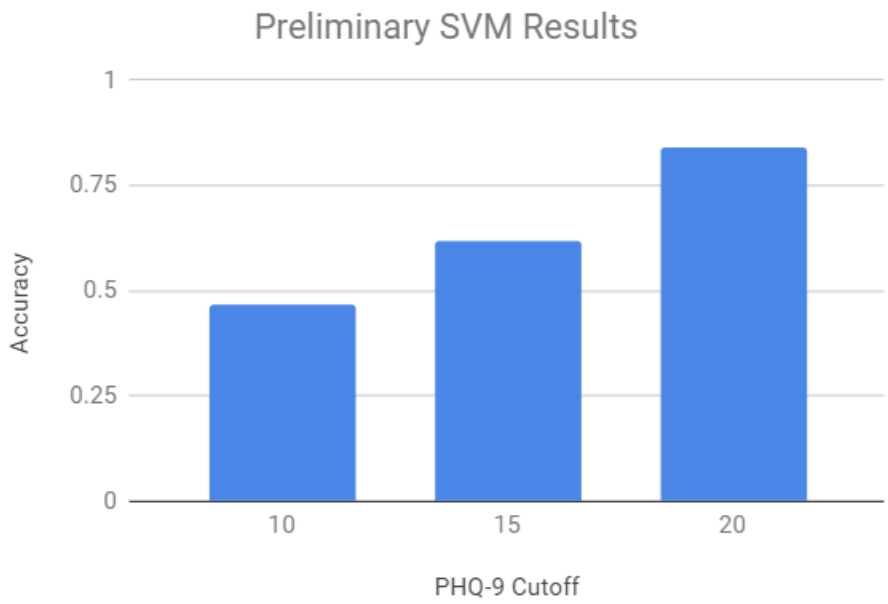

Table 20: Preliminary SVM results

These results were not the final results we've achieved. In a sense, this table provided a benchmark for a certain learner's ability to learn the task at hand. We later present better results in the dataset balancing and feature selection subchapters.

6.3.3. Modality Based Correlation

For modality based correlation, we ran the configuration below through every classifier mentioned under "Different Learners" in the experiments section. For this part, our best performing classifier was, in interesting repetition, the SVM classifier. It is instructive to note that these are results without dataset balancing or feature reduction.

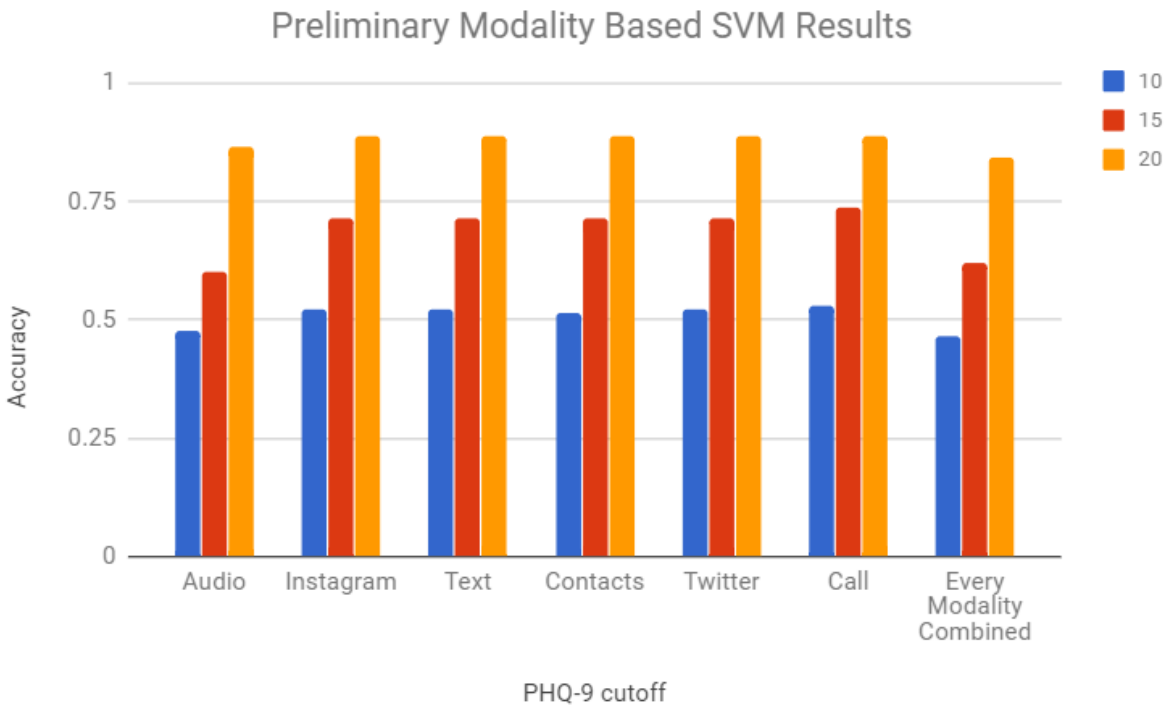

Table 21: Preliminary modality based SVM results on binary labels created with three different PHQ-9 cutoffs

6.3.4. Dataset Balancing

With dataset balancing, we were able to achieve interesting results. For three cutoffs of PHQ-9 scores at 10, 15 and 20, different learners with their hyperparameter configuration and relevant metrics are provided below in Tables 22 through 25.

PHQ-9 Cutoff at 20

SVC (C=14, kernel=”poly”)

Results of two-fold cross validation

label	precision	recall	f1-score	support
0.0	1.00	0.81	0.90	27
1.0	0.84	1.00	0.92	27
avg / total	0.92	0.91	0.91	54

Confusion Matrix: [[22 5]

[0 27]]

Table 22 - Results of two-fold cross validation with PHQ-9 cutoff at 20

PHQ-9 Cutoff at 15

SVC(C=3, kernel=”poly”)

Results of two-fold cross validation:

label	precision	recall	f1-score	support
0.0	0.88	0.84	0.86	76
1.0	0.85	0.88	0.86	76
avg / total	0.86	0.86	0.86	152

Confusion Matrix: [[64 12]

[9 67]]

Table 23 - Results of two-fold cross validation with PHQ-9 cutoff at 15

PHQ-9 Cutoff at 10

SVC(C=10, kernel=”sigmoid”)

Two-fold cross validation

label	precision	recall	f1-score	support
0.0	0.76	0.89	0.82	128
1.0	0.87	0.72	0.79	128
avg / total	0.81	0.80	0.80	256

Confusion Matrix: [[114 14]
[36 92]]

Table 24 - Results of two-fold cross validation with PHQ-9 cutoff at 10

6.3.5. Feature Reduction

With feature reduction in conjunction with dataset balancing, considering the order of magnitude of the number of participants, and the medical significance of the PHQ-9 score cutoff at 10, we decided to concentrate on this particular cutoff and we used feature reduction techniques detailed in the corresponding methods section.

We used this configuration to test out the predictive power of every modality, and the audio modality gave us the most impressive results.

Audio, PHQ-9 cutoff=10

label	precision	recall	f1-score	support
0.0	0.86	0.99	0.92	128
1.0	0.99	0.84	0.91	128
avg / total	0.93	0.92	0.92	256

Confusion Matrix: [[127 1]
[20 108]]

Table 25 - Results of using only audio data with PHQ-9 cutoff at 10

6.3.6. Bagging and Meta-Cost Learning

Before we ran our classifiers on the test set, since our ultimate goal is the generalization of the learners we have experimented with, we put our best performing classifiers through a bagging pipeline. Our bagging method is an ensemble method in which our best performing classifiers for a certain modality are run on different subsets of our dataset with replacement. The bagging estimator then constructs a voting based ensemble meta-estimator that uses multiple instances of our best classifiers that were trained on different random subsets of the dataset, enabling the final meta-estimator to avoid overfitting more that it would have if this process weren't done.

The subset of the features and data points the bagging estimator chooses are treated as hyperparameters of a learner, and are thusly optimized for.

We also optimized our learners so that a recall score for the depressed label is preferred over any other metric. This is dictated by our use case since there is a bigger penalty of not detecting depression than detecting non-depression wrongly. Class weights are adjusted to allow this optimization.

Below you will see results that we have achieved using these bagging estimators for every modality on the training set. Below those results are the results from our test set that we ran upon the conclusion of our experiments with our learners.

We present the training set performances of the final bagging classifiers that use the SVM classifier as their base estimators, with class weights optimized for recall on the depressed label for each. The recall score for the depressed label is emboldened.

We now present the testing set performances of the final bagging classifiers we've trained over the training set. It should be noted that these results were run once at the end of the study, and never again. The recall scores for the depressed label are in bold.

Audio

label	precision	recall	f1-score	support
0.0	0.97	0.78	0.86	73
1.0	0.81	0.97	0.89	72
avg / total	0.89	0.88	0.87	145

Table 26 – Audio with PHQ-9 cutoff at 10 on training

Text

label	precision	recall	f1-score	support
0.0	0.86	0.28	0.42	68
1.0	0.57	0.96	0.71	67
avg / total	0.72	0.61	0.57	135

Table 27 – Text with PHQ-9 cutoff at 10 on training

Twitter

label	precision	recall	f1-score	support
0.0	0	0	0	25
1.0	0.59	1	0.66	24
avg / total	0.24	0.49	0.32	49

Table 28 – Twitter with PHQ-9 cutoff at 10 on training

Call

label	precision	recall	f1-score	support
0.0	0.8	0.17	0.28	70
1.0	0.53	0.96	0.68	69
avg / total	0.67	0.56	0.48	139

Table 29 – Call with PHQ-9 cutoff at 10 on training

GPS

label	precision	recall	f1-score	support
0.0	0.81	0.59	0.68	22
1.0	0.5	0.75	0.6	12
avg / total	0.7	0.65	0.65	34

Table 30 – GPS with PHQ-9 cutoff at 10 on training

Contacts

label	precision	recall	f1-score	support
0.0	0.83	0.07	0.13	74
1.0	0.51	0.99	0.67	73
avg / total	0.67	0.52	0.4	147

Table 31 – Contacts with PHQ-9 cutoff at 10 on training

Instagram

label	precision	recall	f1-score	support
0.0	0.8	1	0.89	4
1.0	1	0.67	0.8	3
avg / total	0.89	0.86	0.85	7

Table 32 – Instagram with PHQ-9 cutoff at 10 on training

We now present the testing set performances of the final bagging classifiers we've trained over the training set. It should be noted that these results were run once at the end of the study, and never again. The recall scores for the depressed label are in bold.

Audio

label	precision	recall	f1-score	support
0.0	0.81	0.59	0.68	22
1.0	0.5	0.75	0.6	12
avg / total	0.7	0.65	0.65	34

Table 33 – Audio with PHQ-9 cutoff at 10 on testing

Text

label	precision	recall	f1-score	support
0.0	0.79	0.41	0.54	27
1.0	0.41	0.79	0.54	14
avg / total	0.66	0.54	0.54	41

Table 34 – Text with PHQ-9 cutoff at 10 on testing

Twitter

label	precision	recall	f1-score	support
0.0	0	0	0	9
1.0	0.31	1	0.47	4
avg / total	0.09	0.31	0.14	15

Table 35 – Twitter with PHQ-9 cutoff at 10 on testing

Call

label	precision	recall	f1-score	support
0.0	0.89	0.11	0.2	70
1.0	0.52	0.99	0.68	69
avg / total	0.71	0.55	0.44	1139

Table 36 – Call with PHQ-9 cutoff at 10 on testing

GPS

label	precision	recall	f1-score	support
0.0	0	0	0	16
1.0	0.36	1	0.53	9
avg / total	0.13	0.36	0.19	25

Table 37 – GPS with PHQ-9 cutoff at 10 on testing

Contacts

label	precision	recall	f1-score	support
0.0	0.64	1	0.78	16
1.0	0	0	0	9
avg / total	0.41	0.64	0.49	25

Table 38 – Contacts with PHQ-9 cutoff at 10 on testing

Instagram

label	precision	recall	f1-score	support
0.0	0	0	0	4
1.0	0.43	1	0.6	3
avg / total	0.18	0.43	0.26	7

Table 39 – Contacts with PHQ-9 cutoff at 10 on testing

As one can observe, the supports in the testing set fall short of providing a meaningful data set in multiple modalities. For these modalities, it seemed to us that the training set results present a more statistically sensible indicator of future promise. We discuss limitations regarding this and suggest future work in the following chapters.

7. Discussion

7.1. Exploratory Willingness to Share Study

Originally, we had planned to focus our development on just the data stored on smartphones, and implement voice and face features in the end if time permitted. However, the results of the exploratory study showed that the population we sampled was much more willing to provide voice and face recordings than all other types of smartphone data. Therefore, we decided to make voice and facial features a higher priority. Instead, we shifted focus away from the data types we found participants were less willing to share, such as browser history, app usage, Facebook posts, and text chat apps such as GroupMe, WhatsApp, and Discord. Some of these data types, such as browser history and app usage, we found were not accessible by Android applications, so we would be unable to gather the data even if participants had been willing to share it. Instead of text chat applications we decided to gather the basic SMS text messages of participants. Even though most participants did not want to share their text messages, we decided to implement the text message gathering because the amount of useful data we thought we might obtain from having pieces of participants' conversation would justify the use of our time. In the end, the data we decided to focus on gathering was Twitter posts, SMS messages, GPS data, call logs, voice recordings, and images of the face.

7.1.1. Limitations

We were given a limited amount of funding for this project, most of which we decided to divert to the data gathering study in order to obtain the best data possible. Therefore, we were very limited in the amount of funding for the exploratory study which resulted in a smaller sample size. We believe the sample size was sufficient for getting the rough idea of the population's willingness to share data. However, for more accurate results we would suggest a large sample size. Another limitation of this study is that participants were not asked whether or not they owned any social media accounts. For example, participants who did not own a Twitter account were still required to answer whether or not they would share their Twitter posts. Given the chance to redo the survey, we would have had the questions specific to social media accounts only be shown to participants who indicated they own the relevant accounts. Furthermore, we

would also add a question asking about how frequently they use the accounts, to see only people who do not use their accounts often feel comfortable sharing their posts.

7.2. Data Gathering Study

The results from the data gathering study were very promising in terms of their quality and quantity. Specifically, the number of participants providing online account data, including Google, Twitter and Instagram was much higher than expected. For example, approximately 28.2% provided Twitter data, and about 21% of US citizens have a Twitter account, so the data acquired provided good exposure to more Twitter users. In terms of our overall objective, the data that was most useful for the machine learning algorithm was primarily related to text messages and audio data. The data gathering study was extremely successful in gathering audio and text data, with 76.5% of the participants providing audio data, and 70.1% providing text data.

7.2.1. Limitations

A phenomenon observed from the results was that willingness to provide data seemed to decrease as the user progressed through the phone app, indicating a "burnout" effect that might influence users' willingness to provide personal data. This might partly explain the lower response rate when it came to online account data provided by the users. A possible solution to this problem would be randomizing the screen order, which would eliminate such biases over the mean. While the data collected from the study was sufficient in training the machine learning algorithms, higher accuracy may be possible had there been more participants, but this was constrained by our limited budget. Running the study on several platforms, and not only Mechanical Turk, could possibly even out some statistical inconsistencies (like the over-representation of Twitter users and under-representation of Instagram users compared to the general populace).

7.3. Machine Learning

The accuracy and precision scores we amassed showed promise in making an app that produces a meaningful prediction of depression. One limitation we have is that our sample population were Amazon MTurk users, and that there is no study done on whether if Amazon MTurk users are a good representation of the United States population at large. This would mean that we do not know for certain if our model would properly generalize for the public. More likely than not, the Amazon MTurk user population would provide data reflective of the population, and we bank on this assumption in conducting our project, but to reiterate, studies must be done to prove that Amazon MTurk users are a non-biased sample of the general United States population.

From our results we've learned that the best predictor of depression in the context of our project is audio features and written language. While the former modality is a non-intuitive predictor whose efficacy in detecting depression is backed by a number of studies, the latter modalities' predictive power presents interesting results, but these results make sense given the interwoven nature of cognition and language. We were able to either come to the vicinity or surpass the results of papers utilizing this modality to detect depression. This comparison is only valid with the assumption that their learners, alongside ours, would perfectly generalize to new data.

The response rates presented in the finding sections made it so that the twitter modality was of no real use. Even with the data we have, this modality presented relatively low predictive power. Future studies should consider gathering more twitter data, or concentrating on this modality to replicate the results that exist in the literature.

The Instagram response rate was relatively moderate, but although this was the case, we couldn't replicate similar results found in the literature concerning the task of correlating Instagram features we used with PHQ-9 scores. This is partly because the data we received often presented challenges that required us to exclude some data from being featurized in order to preserve feature integrity, and this made it so that we had a very small number of data points in the Instagram modality.

7.3.1. Limitations

There are many limitations that concern our machine learning results. The first and foremost is the fact that we did not have enough data for our machine learning algorithms to properly generalize. More data would make it so that our models generalized much better, and would properly function as predictor of depression in the general population. ~350 data points is, from a statistical standpoint, a very small dataset. The limitations that surface from the size of the dataset have reflected themselves in the testing set results.

As one can observe, the number of data points we had for the test set was very low. This was due to the 15% cut of the testing set, in a dataset of 350 participants, most of which provided incomplete data, being a cutoff point that could not get a meaningful number of testing datapoints. More data collection will allow future work in this area to both excel in training learners that generalize well, and produce meaningful testing set results. We believe that for some modalities, testing our classifiers on the testing set was much less of an indicator of generalization than the three fold cross validation done on the training set.

In this vein of logic, when we balance our dataset with the 15 and 20 PHQ-9 cutoffs, which correspond to moderately severe and severe depression, our dataset scales down to the absurd size of 70 and 30 data points respectively. To make any kind of meaningful exploration regarding the detection of severe depression, more data is needed.

In our results we are generally interested in the recall score of the depressed label, since false positives of depression are an inconsequential price to pay if it means that we can detect more true positives of depressed individuals. This made it so that our learners were optimized for the recall score of the depressed label, making our classifiers extremely prone to false negatives.

8. Conclusion and Future Work

In summary, the application we developed has the ability to gather data from the phones of patients entering a hospital, including their SMS text messages, calendar, contacts, files, and call logs, as well as requesting data including a voice recording, Twitter posts, Instagram posts, and Google account data. The application is then capable of using the information gathered in conjunction with the machine learning system we designed in order to give the patient and their doctor an estimate of the severity of depression in the patient. We believe we have created an application that is capable of accurately estimating the level of depression in a patient without requiring the patient to answer any personal questions. The application also does not rely on the honesty of the responses of the patients. We believe that this application could greatly decrease the number of people in the population with undiagnosed depression by making the screening process more accurate as well as making the process faster and easier for the patients.

8.1. Future Work

While powerful, the application could still continue to be improved upon. One major feature that could potentially increase the accuracy of results would be analyzing facial images taken when the patient uses the application. As discussed in section 1.8, analyzing facial features as expression can be used as a powerful tool in detecting depression. The application could have an extra screen that requests permission to take a photo of the face of the user, and then run facial coding on the resulting image. Implementing an extra data type could potentially increase the accuracy of the application. However, implementing facial feature extraction into our application would have taken more time than we had available.

Another area for improvement is in the machine learning systems. While we believe we have created a robust system capable of predicting depression scores and depressed individuals, we only had one and a half quarter (or a little less than a semester) to develop and test the machine learning systems. With more data, a configuration of the machine learning systems which generalize better could be found. If we had more time to run another study and get a larger body of test data, we would have been able to train the machine learning systems on a greater variety of data which could potentially result in a stronger system and better estimates.

In our exploration we encountered promising results in detecting severe depression. We urge future developers of this system to look into an alternative framework of anomaly detection, where the label for depression could be made at a higher PHQ-9 cutoff, allowing the construction of a detection framework that detects severe depression, possibly with greater results.

We also urge the future developer to gather more complete data, where every user provides every modality of data. Per the limitations presented to us with this multi modal dataset where only a small number of participants provided all types of data, and some of this complete data containing questionable data points that needed to be excluded to maintain the integrity of the work done, we couldn't explore multi-modal ensemble methods that used boosting or stacking to achieve better results. Amassing complete data will allow much more meaningful exploration in this avenue, since the groundwork for this multi-modal data set is already constructed, and only needs more participant responses.

9. References

Cohn JF, Kruez TS, Matthews I, et al. Detecting depression from facial actions and vocal prosody. ACII. 2009:1-7.

Matthews I, Baker S. Active appearance models revisited. International Journal of Computer Vision. 2004;60(2):135-164. https://search.proquest.com/docview/1113590423. doi: VISI.0000029666.37597.d3.

Alghowinem S, Goecke R, Wagner M, Epps J, Breakspear M, Parker G. Detecting depression: A comparison between spontaneous and read speech. ICASSP. 2013:7547-7551.

Andreas Strohle. Biological Psychiatry -Review Article: Physical activity, exercise, depression and anxiety disorders. August 23, 2008.

Wang R, Chen F, Chen Z, et al. StudentLife. Proceedings of the 2014 ACM International Joint Conference on pervasive and ubiquitous computing. Sep 13, 2014:3-14.

Ipeirotis P. Analyzing the amazon mechanical turk marketplace. XRDS: Crossroads, The ACM Magazine for Students. 2010;17(2):16-21. http://dl.acm.org/citation.cfm?id=1869094. doi: 10.1145/1869086.1869094.

Prieto VM, Matos S, Álvarez M, Cacheda F, Oliveira JL. Twitter: A good place to detect health conditions. PloS one. 2014;9(1):e86191. http://www.ncbi.nlm.nih.gov/pubmed/24489699. doi: 10.1371/journal.pone.0086191.

De Choudhury M, Counts S, Horvitz E. Predicting postpartum changes in emotion and behavior via social media. Proceedings of the SIGCHI Conference on human factors in computing systems. Apr 27, 2013:3267-3276.

Park M, Cha C, Char M. Depressive moods of users portrayed in twitter. . . https://pdfs.semanticscholar.org/8dd5/8913bd343f4ef23b8437b24e152d3270cdaf.pdf.

Canzian L, Musolesi M. Trajectories of depression. Proceedings of the 2015 ACM International Joint Conference on pervasive and ubiquitous computing. Sep 7, 2015:1293-1304.

Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ. 2016;4:e2537. https://doaj.org/article/d7b5f9efb0964b60adaa4bb861e6e07d. doi: 10.7717/peerj.2537.

Fletcher J. Adolescent depression and adult labor market outcomes. Southern Economic Journal. 2013;80(1):26-49.

Lenert LA, Sherbourne CD, Sugar C, Wells KB. Estimation of utilities for the effects of depression from the SF-12. Med Care. 2000;38(7):763-770.

Rahman I, Humphreys K, Bennet AM, Ingelsson E, Pedersen NL, Magnussen PKE. Clinical depression, antidepressant use and risk of future cardiovascular disease. Eur J Epidemiol. 2013;28(7):589-595.

Current depression among adults united states, 2006 and 2008. Morb Mortal Weekly Rep. 2010;59(38):1229-1235.

Thapar A, Collishaw S, Potter R, Thapar AK. Managing and preventing depression in adolescents. BMJ: British Medical Journal. 2010;340(7740):254-258.

Teixeira CM, Vasconcelos-Raposo J, Fernandes HM, Brustad RJ. Physical activity, depression and anxiety among the elderly. Soc Indicators Res. 2013;113(1):307-318.

Klein DN. Chronic depression: Diagnosis and classification. Current Directions in Psychological Science. 2010;19(2):96-100.

Depression: The treatment and management of depression in adults (updated edition). British Psychological Society; 2010. . https://www.ncbi.nlm.nih.gov/books/NBK63748/.

Y. Yang, C. Fairbairn, J. F. Cohn. Detecting depression severity from vocal prosody. IEEE Transactions on Affective Computing. 2013;4(2):142-150.

Reece, A. G, Danforth C. M, "Instagram photos reveal predictive markers of depression" 2017

Kroenke K, Spitzer R. L, Williams J B Q, J Gen Intern Med. 2001 Sep; 16(9): 606–613. doi: 10.1046/j.1525-1497.2001.016009606.x

Brown, Charlotte et al. "DEPRESSION STIGMA, RACE, AND TREATMENT SEEKING BEHAVIOR AND ATTITUDES." Journal of community psychology 38.3 (2010): 350–368.

Canzian L, Musolesi M. Trajectories of depression. *Proceedings of the 2015 ACM International Joint Conference on pervasive and ubiquitous computing*. Sep 7, 2015:1293-1304.

Cash, Scottye J., and Jeffrey A. Bridge. "Epidemiology of Youth Suicide and Suicidal Behavior." Current opinion in pediatrics 21.5 (2009): 613–619.

Depression: The treatment and management of depression in adults (updated edition). *British Psychological Society; 2010.* . https://www.ncbi.nlm.nih.gov/books/NBK63748/.

Diagnostic and statistical manual of mental disorders: DSM-5 by American Psychiatric Association; American Psychiatric Association. DSM-5 Task Force 2013, 5th ed.

Goodwin, Guy M. “Depression and Associated Physical Diseases and Symptoms.” Dialogues in Clinical Neuroscience 8.2 (2006): 259–265. Print.

Griffiths, Kathleen M et al. “Effectiveness of Programs for Reducing the Stigma Associated with Mental Disorders. A Meta-Analysis of Randomized Controlled Trials.” *World Psychiatry* 13.2 (2014): 161–175.

J. F. Cohn et al., "Detecting depression from facial actions and vocal prosody," 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, 2009.

Paul E. Greenberg et al.,“The Economic Burden of Adults With Major Depressive Disorder in the United States” (2014): 155–162.

Y. Yang, C. Fairbairn, J. F. Cohn. Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing*. 2013;4(2):142-150

Nutt, D. (2008). Sleep disorders as core symptoms of depression. *Dialogues in Clinical Neuroscience* , . doi:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3181883/

Taylor A., Marcus M., Santorini B. (2003) The Penn Treebank: An Overview. In: Abeillé A. (eds) Treebanks. Text, Speech and Language Technology, vol 20. Springer, Dordrecht

Florian, E. (2010). The Munich open Speech and Music Interpretation by Large Space Extraction toolkit. Munich, Germany: Retrieved from https://www.audeering.com/research-and-open-source/files/openSMILE_book_1.0.1.pdf

10. Appendix

Appendix A - Questions of the Patient Health Questionnaire - 9

Over the past 2 weeks, how often have you been bothered by any of the following problems? (Options for each: *Not At all*, *Several Days*, *More Than Half the Days*, or *Nearly Every Day*)

1. Little interest or pleasure in doing things
2. Feeling down, depressed or hopeless
3. Trouble falling asleep, staying asleep, or sleeping too much
4. Feeling tired or having little energy
5. Poor appetite or overeating
6. Feeling bad about yourself - or that you're a failure or have let yourself or your family down
7. Trouble concentrating on things, such as reading the newspaper or watching television
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
9. Thoughts that you would be better off dead or of hurting yourself in some way

Appendix B – Willingness Survey

Introduction: You are being asked to participate in a research study. Before you agree, however, you must be fully informed about the purpose of the study, the procedures to be followed, and any benefits, risks or discomfort that you may experience as a result of your participation. This form presents information about the study so that you may make a fully informed decision regarding your participation.

Purpose of the study: Today as part of emergency room procedure, patients are typically required to fill out mental health questionnaires for ailments such as depression and anxiety. However prior research has found that in place of paper forms, a patient's mental health status can be determined possibly more accurately by analyzing data on their smartphone. This research aims to further advance depression screening and will help develop technology that medical professionals could use to determine patients mental health based on information stored on their smartphones and some other physiological data. This survey investigates the thoughts and opinions about access to different kind of information from your mobile device.

Risks to study participants: There are no physical or mental risks associated with taking this survey. If you feel uncomfortable, you may stop the study at any time. There are also no privacy risks to the subjects because all information gathered will be analyzed in aggregate form and this is an anonymous survey so no subject will be identified. **To be clear, this survey only asks how willing the participant is to share the information and does not actually gather any of the listed information.**

Benefits to research participants and others: The information learned from this study could aid in

the early detection of mental health issues and facilitate a medical professional's ability to intervene.
Record keeping and confidentiality: Records of your participation in this study will be held confidential so far as permitted by law. However, the study investigators, and under certain circumstances, the Worcester Polytechnic Institute Institutional Review Board (WPI IRB) will be able to inspect and have access to confidential data that identify you by name. Any publication or presentation of the data will not identify you.

Q13 Are you 18 and above?

- ○ Yes (1)
- ○ No (2)

Q11 What is your gender?

- ○ Male (1)
- ○ Female (2)
- ○ Others (3) _____
- ○ Prefer not to answer (4)

Q12 What is your age range?

- ○ 18 - 25 (1)
- ○ 26 - 35 (2)
- ○ 36 - 45 (3)
- ○ 46 - 55 (4)
- ○ Above 55 (5)

Q14 Please select the option that best describes your current employment status.

- ○ Student (1)
- ○ Employed (2)
- ○ Unemployed (3)
- ○ Retired (4)
- ○ Other (5) _____

Q15 What is your major or area of concentration?

_____

Q10 Have you ever been treated in an emergency clinic?

- ○ Yes (1)
- ○ No (2)

Q1 Social Media:

How willing would you be to allow a member of the medical staff to run software that retrieves

the following information from your social media accounts and feeds it into a program that estimates your mental health status:

	Completely Unwilling (1)	Somewhat Unwilling (2)	Unsure (3)	Somewhat Willing (4)	Completely Willing (5)
Your tweets on Twitter (1)	○	○	○	○	○
Your Twitter username (2)	○	○	○	○	○
Access to posts on your Facebook (3)	○	○	○	○	○
Your messages on GroupMe, Discord, WhatsApp, etc (4)	○	○	○	○	○

Q3 Phone Data:

How willing would you be to allow a member of the medical staff to run software that retrieves the following information stored on your phone and feeds it into a program that estimates your mental health status:

	Completely Unwilling (1)	Somewhat Unwilling (2)	Unsure (3)	Somewhat Willing (4)	Completely Willing (5)
Your phone's GPS data (1)	○	○	○	○	○
Your phone's gyroscope/accelerometer data (2)	○	○	○	○	○
Your phone's browser history (3)	○	○	○	○	○
Your phone's call logs (4)	○	○	○	○	○
Your phone's app usage data (which Apps are open and how long) (5)	○	○	○	○	○

Q4 Physiological Data:

How comfortable are you performing the following actions when prompted to by a doctor or other member of the medical staff?

	Extremely uncomfortable (1)	Somewhat uncomfortable (2)	Neither comfortable nor uncomfortable (3)	Somewhat comfortable (4)	Extremely comfortable (5)
Speaking a phrase into a microphone (1)	○	○	○	○	○
Permitting an image of your face to be captured and analyzed (2)	○	○	○	○	○

Q16 Please select the age range that you fall under.

- ○ 50 or older (1)
- ○ 41 - 50 (2)
- ○ 31 - 40 (3)
- ○ 21 - 30 (4)
- ○ under 20 (5)

Q6 If you have any comments that you would like to share please put them here. Your help is much appreciated!

Appendix C: Preliminary Accuracy Scores of 27 Class Classification Task between Every PHQ-9 Question and Every Modality using SVM

	Audio	**Instagram**	**Text**	**Contacts**	**Twitter**	**Call**	**Every Modality**
1	0.298643	0.343891	0.343891	0.334842	0.343891	0.343891	0.294118
2	0.235294	0.357466	0.357466	0.357466	0.357466	0.375566	0.230769
3	0.257919	0.294118	0.294118	0.289593	0.294118	0.280543	0.266968
4	0.285068	0.334842	0.334842	0.307692	0.334842	0.321267	0.262443
5	0.303167	0.330317	0.330317	0.312217	0.330317	0.298643	0.307692
6	0.280543	0.298643	0.298643	0.303167	0.298643	0.289593	0.20362
7	0.280543	0.357466	0.357466	0.339367	0.357466	0.375566	0.271493
8	0.533937	0.59276	0.59276	0.59276	0.59276	0.588235	0.488688
9	0.502262	0.561086	0.561086	0.547511	0.561086	0.556561	0.493213
PHQ-9 Score Sum	0.040724	0.063348	0.063348	0.063348	0.063348	0.067873	0.049774

Appendix D: Preliminary Accuracy Scores of 27 Class Classification Task between Every PHQ-9 Question and Every Modality using Logistic Regression

	Audio	Instagram	Text	Contacts	Twitter	Call	Every Modality
1	0.285068	0.339367	0.339367	0.339367	0.339367	0.352941	0.294118
2	0.276018	0.361991	0.361991	0.361991	0.361991	0.366516	0.266968
3	0.271493	0.298643	0.298643	0.298643	0.298643	0.294118	0.289593
4	0.307692	0.321267	0.321267	0.321267	0.321267	0.321267	0.298643
5	0.257919	0.303167	0.303167	0.303167	0.303167	0.307692	0.248869
6	0.352941	0.325792	0.325792	0.325792	0.325792	0.325792	0.343891
7	0.307692	0.366516	0.366516	0.366516	0.366516	0.366516	0.298643
8	0.588235	0.58371	0.58371	0.58371	0.58371	0.58371	0.588235
9	0.542986	0.547511	0.547511	0.547511	0.547511	0.547511	0.542986
PHQ-9 Score Sum	0.090498	0.076923	0.076923	0.076923	0.076923	0.076923	0.090498