2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)

Objective Assessment of Depressive Symptoms with Machine Learning and Wearable Sensors Data

Asma Ghandeharioun[†], Szymon Fedor[†], Lisa Sangermano[*], Dawn Ionescu[*],
Jonathan Alpert[*], Chelsea Dale[*], David Sontag[‡], and Rosalind Picard[†]
[†] *Media Lab, MIT, Cambridge, MA 02139*
Email: {asma_gh, sfedor}@mit.edu, picard@media.mit.edu
[*]*DCRP, MGH, Boston, MA 02114*
Email: {lsangermano, dionescu, jalpert, cfdale}@mgh.harvard.edu
[‡]*CSAIL, MIT, Cambridge, MA 02139*
Email: dsontag@mit.edu

Abstract—**Depression is the major cause of years lived in disability world-wide; however, its diagnosis and tracking methods still rely mainly on assessing self-reported depressive symptoms, methods that originated more than fifty years ago. These methods, which usually involve filling out surveys or engaging in face-to-face interviews, provide limited accuracy and reliability and are costly to track and scale. In this paper, we develop and test the efficacy of machine learning techniques applied to objective data captured passively and continuously from E4 wearable wristbands and from sensors in an Android phone for predicting the Hamilton Depression Rating Scale (HDRS). Input data include electrodermal activity (EDA), sleep behavior, motion, phone-based communication, location changes, and phone usage patterns. We introduce our feature generation and transformation process, imputing missing clinical scores from self-reported measures, and predicting depression severity from continuous sensor measurements. While HDRS ranges between 0 and 52, we were able to impute it with 2.8 RMSE and predict it with 4.5 RMSE which are low relative errors. Analyzing the features and their relation to depressive symptoms, we found that poor mental health was accompanied by more irregular sleep, less motion, fewer incoming messages, less variability in location patterns, and higher asymmetry of EDA between the right and the left wrists.**

1. Introduction

Depression is the leading cause of ill health and disability worldwide: According to the latest estimates from WHO, more than 300 million people are now living with depression, an increase of more than 18% between 2005 and 2015 [1]. Historically, diagnosing and tracking depressive symptoms has been accomplished through periodic assessment with structured or unstructured clinical interviews using standardized symptom rating scales. This approach, which was invented in the 1960s, is based largely on subjective self-report, and has limited utility in fully characterizing clinically meaningful subtypes of depression. Also, this current "descriptive" way of diagnosing depression is limited in its ability to predict the course of illness or to capture variations of the disease over days.

An important paradigm shift is happening today: Psychiatry and the clinical neurosciences are moving from relatively narrow neurochemical models of disease, based on inferences about the pharmacological mechanisms of available psychotropic medications, to broader anatomical and neurophysiological understanding of emotion, behavior, cognition and their disorders [2]. This shift is important, not only because it provides a new understanding of the neuroscientific basis of psychiatric disorders, but also because it leads to the development of novel strategies for diagnosis and assessment. Researchers are increasingly developing objective mobile data-driven biomarkers for many healthcare conditions, including depression (e.g. [3]). We anticipate that the development of reliable biomarkers will help improve the diagnosis and assessment of depression, prediction of treatment response, and early detection of response, remission and relapse. To date, there is no set of reliable biomarkers to assess depression.

In this paper, we advance the state of the art in the development of biomarkers by providing a new way, based on passive sensing, to estimate depressive symptoms as measured by the Hamilton Depression Rating Scale (HDRS). The method utilizes data from E4 wearable sensors [4] and embedded sensors within an Android phone. Experience sampling, continuously capturing self-reported depressive symptoms, can be overwhelming for a patient in the long-run. Being able to estimate HDRS scores accurately using passive data could potentially improve the scalability of depression prognostication as well as its objectivity. In the meantime, it enables a fertile ground of research for providing timely interventions to individuals who show signs of relapse. Also, we believe that there is more value in a regression analysis as opposed to a classification between different severity levels of depressive symptoms. With regression, we may obtain a more accurate and precise understanding of the progression of the disease.

In our dataset, HDRS has been captured bi-weekly by a

978-1-5386-0563-9/17/$31.00 ©2017 IEEE

clinician, as part of their standard practice. Thus, we utilize a two-step prediction process: First, we use a surrogate (self-reported data) to predict HDRS and in doing so, impute the missing HDRS values (from the dates when the HDRS was not assessed by a clinician) to construct an increased dataset "HDRS-I". Second, we use the passive phone and wearable sensor measures for predicting the HDRS-I values.

2. Background and Related Work

Over the past decade, affective computing researchers have utilized wearable sensors and phone usage patterns to detect stress, happiness, and mental wellbeing (e.g. [5], [6]). We hypothesize that similar underlying phenomena quantifying mood can help assess mood disorders as well.

Numerous researchers have demonstrated the use of mobile-based Experience Sampling Methods to monitor people's depression, e.g. [7], [8]. In these studies, the depressed patients are asked to fill out regular surveys about mood, behavior, sleep etc. on their mobile phones. The self-reports have several limitations. They can be unreliable as the response rate may depend on the current mood of the patients. Moreover, they are subjective since such logs are recorded by the patients themselves and the answers may vary with factors including mood, weather, social-demands, or patient's memory. Finally, frequently answering the mobile surveys is cumbersome, which may introduce bias or result in reduced adherence.

Several studies have proposed to measure passively objective parameters in controlled environments (hospital or laboratory). One of the first efforts to assess how long-term physiology and behavior of individuals are correlated with changes in depression was the LiveNet project [9]. The LiveNet platform, which monitored skin conductance, heart rate, activity and voice, was evaluated on six psychiatric inpatients. More recently, Valenza et al. [10] demonstrated the use of electrocardiogram and respiration signals collected in a hospital to assess depression. Although these studies show promising results, we aim at a harder problem: to continuously and unobtrusively monitor people during daily life in order to identify possible biomarkers of depression.

The MONARCA project [11], which developed tools for assessment and prediction of mood episodes in bipolar disorder, focused on analytics tools and validating them with a group of 20 patients. Also, scholars have studied phone usage correlates of mental health and depressive symptoms (e.g [12], [13]). Other researchers have looked at audio/visual cues including facial expressions, head movement, vocalization, and vowel production to predict depression severity (e.g. [14], [15], [16]). However, many of these studies have been validated based on self-reported standard depression scales, like Patient Health Questionnaire (PHQ-9) [17], rather than on clinical measurements. In this paper, we aim to fill the gap by including clinical assessment of depressive symptoms using Hamilton Depression Rating Scale (HDRS) as scored by the expert clinician in a patient interview. The clinical form of HDRS data is collected in a face-to-face meeting bi-weekly as it has been demonstrated that intensive assessment of depression may have a positive impact on the assessment score [18]. We then impute the depression level of the remaining dates using Machine Learning that incorporates daily patient self-reports.

Most previous work has addressed a classification problem, usually binary, within this area [19], [20]. Some captured only categorical label variables, while others transformed an inherent regression problem into a classification problem, and in doing so relaxed the problem; for example, they only included the highest and lowest values of the depression range and did not address the "middle". However, depressive symptoms change continuously and which way they are shifting is important. To better understand and prevent worsening of depression, it is not enough to distinguish between extremely severe and extremely mild depressive symptoms: We aim to measure progressive change of symptoms in order to enable just-in-time interventions before depression becomes severe.

3. Study Protocol

Patients diagnosed with MDD from Massachusetts (n=12) completed an 8-week protocol. Participants included 9 females and 3 males from white, hispanic, african-american, and asian races and aged between 20 and 73 years old (mean=37, std=17). The protocol involved tracking depressive symptoms and mobile phone usage. Movisens [21] was used to measure incoming and outgoing text messages and phone calls, location, app usage, and screen on/off behavior. Patients also wore Empatica E4 wristbands [4] that recorded accelerometer data and electrodermal activity 23 hours a day. Measurements were processed to obtain daily aggregate measures. Participants were clinically assessed for depression symptoms biweekly using the HDRS. Tab. 1 summarizes the number of observations for each modality. In the next section, we explain the detailed measurements in each modality and the feature generation.

TABLE 1: Dataset summary after computing daily features.

Modality	# of Datapoints
Physiological signals	540
Phone passive usage data	605
Interactive surveys	503
Clinical measures	59

4. Feature Architecture

4.1. Physiological Signals

E4 sensors worn on each wrist captured continuous electrodermal activity (EDA) via the measurement of skin conductance (4Hz sampling rate), temperature (4Hz sampling rate), and 3-axis accelerometer data (32 Hz sampling rate). In order to better understand the user's behavior within the day, we introduce 6-hour intervals, labeled as morning, afternoon, evening, and night. The 6-hour interval provides a balance between granularity and ratio of missing values.

We also calculate aggregate daily measures. Any feature explained below has been calculated for all these intervals.

We first filtered out the EDA signal when the corresponding skin temperature was below 31°C to exclude the measurements when the sensor was not worn. Then we applied the 6th order Butterworth low-pass filter (1Hz cutoff frequency). We calculated mean EDA and the fraction of time the sensor was recording the signal. We also computed the number of skin conductance response (SCR) peaks and their average amplitude using the method from Gamboa [22]. There are indications that skin conductance level may distinguish between depressed and healthy individuals [23]. Also, previous research has shown that asymmetry in EDA between the wrists can provide extra affective information [24]. Thus, we also encoded asymmetry in different ways: the difference between average EDA value, difference between number of SCRs, and difference between SCL and SCR signals using Convex Optimization Approach [25].

We applied the the 5th order Butterworth low-pass filter (10Hz cutoff frequency) to the accelerometer data. We then translated the output into motion features by calculating the vector magnitude, VM of the z-axis acceleration data using the following formula:

$$VM = \sum_{t=0}^{N} VM_t + |R_{(z,t)} - M_z| \tag{1}$$

where $R_{(z,t)}$ is the raw accelerometer z-axis sample, M_z is the running mean in a 5-second window of the z-axis signal, and N is the number of raw data samples received in one second.

Next, we calculated average, median, and standard deviation of motion for the mentioned time intervals as well as the fraction of time in motion. We also kept meta-data such as the fraction of time within the time interval that the data were not missing.

We calculated objective sleep based on accelerometer data for 30 second epochs using the ESS method described in [26]. We calculated sleep duration, sleep onset time (time elapsed since noon), maximum duration of uninterrupted sleep, number of wake-ups during the night, and the time of waking up (time elapsed since midnight). We also computed a sleep regularity index (SRI):

$$SRI = \frac{1 + \frac{1}{T - \tau} \int_0^{T\tau} s(t)s(t+\tau)dt}{2} \tag{2}$$

where data were collected for $y = [0, T]$, $\tau = 24$, $s(t) = 1$ during sleep and $s(t) = -1$ during wake. The SRI ranges between 0 (highly irregular sleep) and 1 (consistent sleep every night). We also included meta-data such as the fraction of time that data were being recorded over nighttime (between 8pm-9am) as well as over the period of 24 hours.

4.2. Phone Passive Usage Data

We utilized Movisens [21] on Android to collect measures of how the participant is using his or her mobile phone and how s/he is interacting with other people using the mobile phone. More specifically, we captured meta-data of calls, text messages, app usage, display on/off behavior, and location. Passive data were captured 24/7. The content of the calls/texts, actual phone numbers, websites visited, and the content of the applications were not collected.

Following the steps of previous researchers in generating features from passive phone data [5], we introduce 3-hour intervals in order to better understand the user's daytime behavior. For example, [6am-9am] represents early morning while [9pm-12am] corresponds to late evening. We also calculate aggregate daily measures.

For quantifying call data, we calculate the number of incoming, outgoing, and missed calls daily and over the 3-hour periods within the day. In a similar manner, we calculate mean, median, and standard deviation (SD) of the duration of incoming, and outgoing calls. Finally, we calculate the incoming/outgoing ratio both for the number of calls and the duration of calls on a daily basis.

For quantifying SMS data, we use a similar approach, we calculate the number of incoming and outgoing texts daily and over 3-hour periods within the day. We also calculate a daily incoming/outgoing ratio of the number of text messages received or sent respectively.

Turning the display on/off is also an indication of phone usage. Thus, we look at the mean, median, and SD of duration of screen on within the mentioned intervals. We also calculate the number of the times the screen has been turned on over these periods. Note that these two correspond to different behaviors; Long screen-on duration is related to actively using the phone while a great number of screen-ons is related to consistently checking the phone which might be a sign of anxiety or anticipation.

For location data, we calculate mean, median, and SD of latitude and longitude along with the number of data points that have been captured for each time period. We calculate total location mean, median, and SD by averaging values from latitude and longitude.

For app usage, we encode the app category using the following list: game, email, web, calendar, communication, social, maps, video streaming, photo, shopping, and clock. Then, we calculate the total duration and the number of app category usage in the different mentioned time intervals.

4.3. Interactive Surveys

Using the Movisens [21] on the mobile phone, we administer short questionnaires about overall health condition, sleep, mood, stress, anxiety, alcohol/drugs/caffeine usage, and social interaction; these should be completed each day upon awakening, at bedtime and twice during the day at random times, during the entire length of the study. For assessing mood, we have used Positive and Negative Affect Schedule (PANAS) [27], one of the most prevalently used scales for measuring affect. The 20 item questionnaire has been splitted into two 10-item questions that were administered twice during the day at random times.

First, we preprocessed the data: we added how long it took the participant to fill in the survey and removed responses that took less than a second and are likely noise. This meta-data can also be informative; for example, long pauses while responding to surveys may represent motor slowing (a common symptom of depression), cognitive load, trouble remembering, or not being sure about the response. Short response time, on the other hand, may represent trivial answers or not reading through the questions. We calculate total alcohol (standard drink measure) and caffeine consumption (milligram) by summing the relevant features from the survey. We convert categorical features to their one-hot representation. We include day of the week as it has been shown to influence the aggregate number of smiles which can be an indication of positive valence mood [28].

Since HDRS is closely related to self-reported mood, we add more detailed mood information. First, we calculate total positive affect (PA) and negative affect (NA) on a daily basis by averaging responses to relevant survey questions. We add an average of the past week's PA and NA. We add a weighted average of PA and NA, when the effect of affect diminishes exponentially overtime when going back in history, e.g., yesterday's mood is half as important as today's mood in the weighted average measure. We calculate the NA/PA ratio for the daily, average weekly, and weighted average weekly measures. To capture mood oscillation, we include the standard deviation of mood on a weekly basis and for the duration of the study.

4.4. Clinical Measures

During each biweekly visit, participants are assessed by the clinician for depressive symptoms using the HDRS. HDRS is a standard test for quantifying depressive symptoms which ranges between 0 and 52. Tab. 2 summarizes the depression severity in relation to HDRS.

TABLE 2: HDRS values and levels of depression severity.

HDRS	Depression Severity
0-7	Normal
8-13	Mild Depression
14-18	Moderate Depression
19-22	Severe Depression
≥23	Very Severe Depression

5. Models

5.1. Feature Transformation and Selection

Combining the carefully-crafted features results in over 700 features for our dataset. Compared to the small number of data points we have, this number of features can easily result in over-fitting the model to the training set. One possibility is to use regularization tricks such as L1 to enforce selection of only a small number of features. However, for features that are non-linearly related, transforming the features into a new space through a non-linear transformation can be more beneficial. For example, several noisy measurements of a similar phenomena may not be informative on their own, but a transformed version of them can be a better predictor. Toward this end, we tested PCA, kernel PCA with radial-basis function kernel, and truncated SVD methods to reduce the dimensionality of our feature-set. We bound the number of selected features while keeping as few features as possible to explain the variance of data.

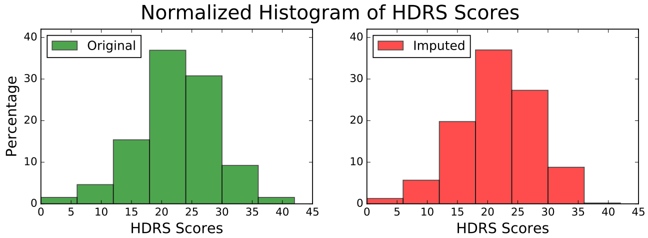


Figure 1: Normalized histogram of HDRS scores before and after imputation.

We created 3 datasets: one including all features, one including daily features only, and one including the daily features and the features of the previous day. We conducted the feature transformations on these three datasets.

5.2. HDRS Imputation Based on Survey Data

Studies have confirmed relationships between self-reported affect and clinical ratings of depression (e.g. [29]). In our dataset, we see a strong correlation between average weekly negative/positive affect $(M = 0.86, SD = 0.38)$ and HDRS scores $(M = 19.64, SD = 7.60)$, $r = 0.70, p = 0.00, n = 44$[1]. This observation suggests utilizing self-reported survey data to estimate the gold-standard measure, HDRS-I, in between clinical assessments. The input features included daily PA, NA, and NA/PA ratio. We have also included average and standard deviation of these values over the past week and over the whole period of the study for the patient. Also, weekly weighted average of these values have been included where the effect of affect diminishes exponentially over time. We then impute the missing values to construct a 10-times-larger dataset. Hence, we use two sets of models to predict the HDRS score from survey data: regularized regression and robust-to-outlier methods.

5.2.1. Regression Models. The regression methods include lasso, ridge, and elasticNet which use L1, L2, and a combination of the two as regularization metrics, respectively. Note that the L1 regularization term acts as a feature selection mechanism by pushing coefficients of most of the variables to be exactly zero, while L2 pushes many coefficients to near zero values but does not remove them completely. We also included regression without regularization with the reduced and transformed features.

[1] Data points with missing mood reports from surveys have been removed from this analysis. This reduces the number of data points from 59 available HDRS measurements to 44.

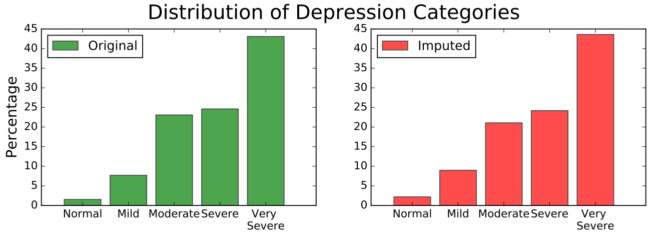


Figure 2: Distribution of depression categories before and after imputation.

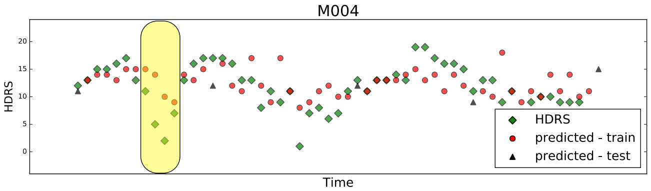


Figure 3: Time-series of HDRS scores (original and predicted) for one sample user over the course of 8 weeks.

5.2.2. Robust Models. To be robust against outliers or errors in formulation of the model, we include Theil-Sen estimator, random sample consensus (RANSAC), and huber algorithms. These models have a built-in sampling procedure that allows a fraction of data points to be outliers.

5.2.3. Validation. For validation, we split the data into 90% training and 10% testing. We use leave-one-out cross-validation on the training set to select the best model and use it for imputing missing HDRS values.

5.3. HDRS Prediction Based on Sensor Data

After imputing HDRS scores, the new dataset HDRS-I is over 500 points. This dataset is still not large enough to be able benefit from state-of-the-art neural network techniques [2]. Thus, we focus on models that do not require enormous amounts of training data. Note that self-reported affect measures have been used only in the imputation phase and are excluded from this step. The HDRS prediction phase solely uses the passive wearable and phone sensor data.

5.3.1. Regression Models. Similar to the imputation phase, we use lasso, ridge, elasticNet, and unregularized regression.

5.3.2. Robust Models. Similar to the imputation phase, we use Theil-Sen, RANSAC, and huber methods. However, we loop through a larger list of parameters to optimize within each model. We should note that these models do poorly when the feature set is large. Thus, we only use them for the subsets or the reduced version of the data.

5.3.3. Boosting. Boosting combines week regressors sequentially to improve performance. We use adaptive boosting (AdaBoost) and Gaussian boosting in this category.

5.3.4. Random Forest. Random Forest is an ensemble method with multiple decision trees. We use random forest with different numbers of estimators.

5.3.5. Gaussian Process. Since natural phenomena usually follow a Gaussian distribution, we use Gaussian Process with different regularization parameters and different numbers of restart points to model the data.

5.3.6. Customized Ensemble Method. Finally, we combine the results from these different regressors to get a more robust estimator. The ensemble method first finds a set of k nearest neighbors from the training set for each point. It then chooses the model that performs best on that set as the estimator for this point. The heuristic behind this method is that slight modifications in the feature set do not change the output drastically. Thus, if a classifier is working well on similar points, chances are it works well for the current point, as well. Looking at k nearest points as opposed to only the most similar point is for smoothing purposes. Note that as the points become higher dimensional, the distance between them becomes less meaningful in explaining similarity between the points. Thus, we only use the first 5 reduced features based on kernel-PCA and create a KD tree and find the k nearest neighbors to the point at hand.

5.3.7. Validation. In real life, some depressed patients see a doctor and get clinical assessments at some point in their life. One major issue is a high relapse ratio and not being able to regularly visit the doctor to re-assess the improvement or worsening of depressive symptoms. In such cases, our method could be easily deployed in real life to passively monitor the patients after the diagnosis. Thus, we will assume that we have at least some history for each user.

For validation, we split HDRS-I dataset into 90% training and 10% hold-out testing. However, the test set is only chosen from the original HDRS values (rather than the imputed ones). We further choose the test set in a way to mimic the real-life deployment scenario: no data point from the first two weeks is selected as test data. We use 10-fold-cross-validation on the training set to select the best model and use it for predicting HDRS values.

6. Results and Discussion

6.1. Imputation Phase

Root mean squared error (RMSE) is the primary metric used to validate the imputation phase. Table 4 shows the selected best model based on having the lowest RMSE on the validation set. Then we report the RMSE on the hold-out test set for each model. This model is ridge regression on the subset of mood features from the survey data, obtaining

[2] For example, long short-term memory (LSTM) network, a strong model that retains temporal information, performs as well as predicting the average value. We have ran LSTM on the dataset as well as an augmented version of it. For augmentation, we have added $x * 0.01 * SD_f$ to each feature f where x is a random number between -0.5 and 0.5 and SD_f is the standard deviation of the values for that feature.

TABLE 3: Best prediction model.

Model Type	Model	Parameters	Dataset	RMSE Validation	 Test	Baseline Average	 Median
Regression	Regression		Kernel PCA subset	5.2	4.9	7.1	7.1
Robust	Ransac	ms=0.3	Kernel PCA subset	5.0	4.9	7.1	7.1
Boosting	AdaBoost	n=50, lr=1	Subset data	5.5	4.6	7.1	7.1
Random Forest	-	n=15	Subset data	5.4	4.6	7.1	7.1
Gaussian Process	-	α=0.1, n=5	Kernel PCA subset	5.3	5.5	7.1	7.1
Overall Ensemble		k=1	selected by individual models	5.8	4.5	7.1	7.1

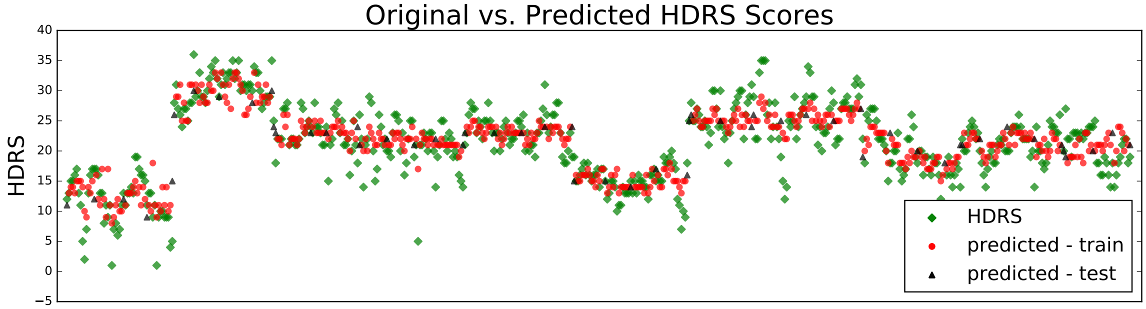
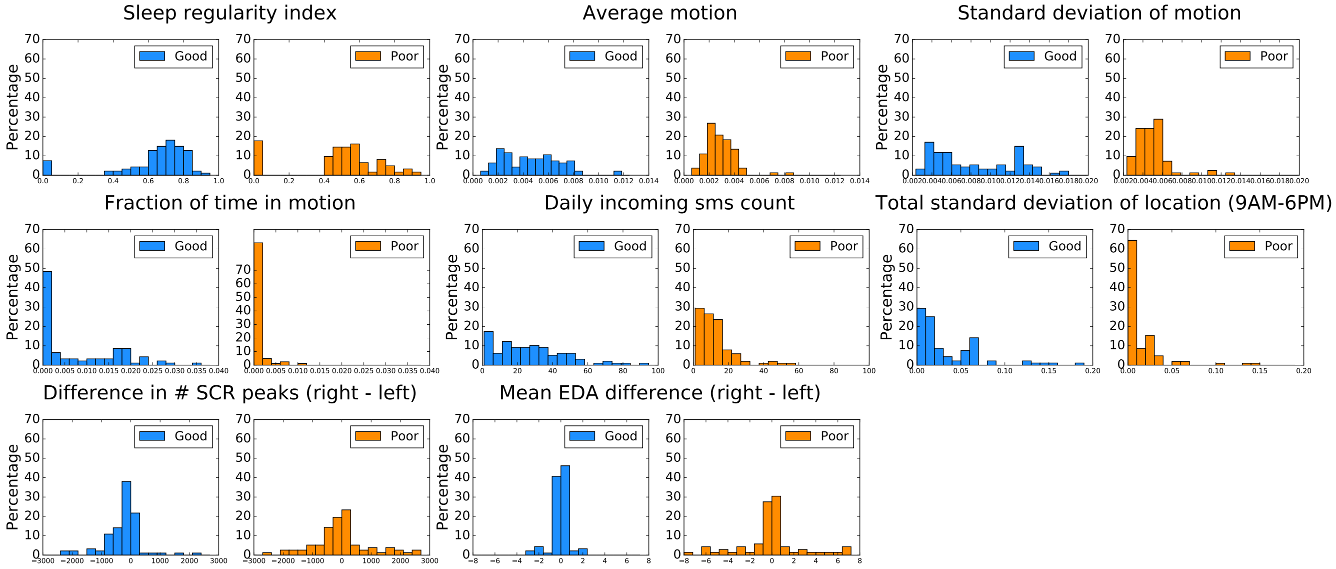


Figure 4: Original and predicted HDRS scores for daily data from all patients over the course of 8 weeks.

Figure 5: Distribution of features that are significantly different between days with good vs. poor mental health.



a test RMSE of 2.8. A baseline prediction of reporting the average or median HDRS score results in an RMSE of 6.8.

Looking more closely at the model provides insights about how the mood features correspond to the HDRS score. Consider the coefficients with the highest absolute values: The coefficient for weekly average positive affect is -9.3, confirming that reported positive affect is negatively associated with HDRS score. Another interesting observation is the -7.4 coefficient of standard deviation of positive mood in the previous week. Depression is usually accompanied by anhedonia, withdrawal, and loss of engagement, which result in consistent low positive mood. Thus, a normal variation in positive mood is negatively associated with HDRS score. At the same time, we see positive association between the average weekly negative affect and the HDRS score, shown by a positive 2.8 coefficient.

To further test the validity of the imputation model, we plotted the distribution of HDRS scores before and after the imputation (Fig. 1), and we used the Kolmogorov-Smirnov (KS) test to compare these two distributions. KS could not reject the null hypothesis of samples coming from a common distribution. [3] Moreover, we examined the predicted levels (based on Table 2) of depression severity before and after

[3] $D_{original}(M = 21.5, SD = 6.4), D_{imputed}(M = 21.2, SD = 6.3); ks - statistic = 0.08, p = 0.83$. The small ks-statistics and large p-value show that we cannot reject the null hypothesis.

imputation. Fig. 2 shows the bar chart of the distribution of depression severity categories. We also tested these two discrete-valued distributions and found they were not significantly different[4].

6.2. Prediction Phase

We primarily validate the new prediction model using RMSE. Table 3 shows the best performing model in each category and the overall customized ensemble method. The test RMSE for the ensemble method is 4.5 while it is 7.1 for the average or median baseline prediction.

To provide understanding of the predictions, we have visualized the time-series of HDRS-I values for a sample user (Fig. 3). Each point represents the HDRS-I value for a day. Green diamonds shows original values (either through clinical assessment or imputation). Red circles and gray triangles show the predictions for train and test points respectively. One interesting observation about this plot is the large prediction error in the highlighted area. A clinician we work with suggested it might be due to the placebo effect of being in the study. Many patients begin to feel better soon after joining the study, and report this, but they fall back into their depressed trend after the novelty effect wears off. We hypothesize that the placebo effect influences momentary assessment of mood quickly, while it is not adequate to influence behavioral or physiological signals. Thus, we see the red dots showing that the prediction based on the objective passive data, while it improves a little, does not improve as much as the self-reported (or their imputed) values. Fig. 4 visualizes the predicted and original values for all the data points from different users with the same color coding. As shown in both figures, the predictions follow the overall trend very well but miss the short term variations of HDRS-I. We should note that HDRS is meant to measure depressive symptoms over the course of two weeks. Thus, from a clinical perspective, it is not supposed to vary much over consecutive days.

The final prediction based on the ensemble algorithm is a combination of different methods and sometimes non-linear feature transformations of the "subset data". To gain deeper understanding of the relationship between the feature space and the resulting predictions we create two classes of points: the top 20% and the bottom 20% of the predicted HDRS-I scores. The former group represents days when the patient is doing very poorly and the latter represents the days when the patient is doing well or showing minimal depressive symptoms. We have compared the distribution of all the features from the "subset data" for these two groups using the KS test. Table 5 summarizes the 8 most significantly different distributions (highest ks-statistics and lowest p-values) and Fig. 5 depicts the differences where blue and orange show the good and poor mental health group respectively. The poor mental health group has more irregular sleep, moves much less on average, shows less motion variability, and is active a lower percentage of the time. Also, this group receives fewer incoming messages and has less variable location patterns. Another interesting finding is the EDA asymmetry. The number of skin conductance responses (SCR) between left and right wrist are mostly similar in the good mental health group. However, we see stronger asymmetry (more SCR peaks on the right wrist) for the poor mental health group. A similar trend is observed in average EDA magnitude.

TABLE 4: Best imputation model.

Model Info.	Name Dataset	Ridge (L2-Regularized Regression) Mood Subset (PANAS)
RMSE	**Validation**	3.4
	Test	2.8
	Baseline 1 (Average)	6.8
	Baseline 2 (Median)	6.8

TABLE 5: Most significantly different distributions of feature values for days with good vs. poor mental health.

Category	**Feature**	**ks-statistic**	**p-value**
Sleep	Sleep regularity index	0.51	$2e-9$
Motion	Average motion	0.49	$3e-10$
	SD of motion	0.47	$3e-9$
	Fraction of time in motion	0.44	$4e-8$
Communication	Daily # incoming SMS	0.44	$3e-9$
Location	Total SD of location (9AM-6PM)	0.34	$8e-6$
Physiology	Difference in #SCR peaks (right-left)	0.29	$8e-4$
	Mean EDA difference (right-left)	0.21	$4e-2$

These analyses are based on data from 12 participants from Massachusetts. Further studies are needed to confirm if the findings are generalizable to other populations, as well.

6.3. Limitations

In the future, we would like to explore the feasibility of our methods for other scenarios, for example having hold-out test subjects resembling when some patients have no observed data in the training phase. We know that there is some interdependency within patients. The variety of our prediction models and the ensemble methods can learn to account for individual differences. However, for future study, we would like to explicitly model that by comparing against mixed-effect models and modeling the patients' variations from their baseline. In this paper, we included several post-hoc analyses to discover more informative data streams. For future study, we would like to further explore the discriminative power of those features separately.

7. Conclusion

In this paper, we showed the feasibility of continuously measuring depressive symptoms using a new method that requires only passive data captured from built-in sensors of a regular android phone and E4 wristbands, including measures of EDA, sleep patterns, motion, communication, location changes, and phone usage patterns. Using a novel combination of machine learning techniques, we were able

[4] $ks - statistic = 0.01, p = 1.00$

to predict the imputed Hamilton Depression Rating Scale (HDRS) values on a hold-out set, obtaining a low error rate of 4.5 RMSE. Moreover, a post-hoc statistical analysis showed that poor mental health was associated with more irregular sleep, less motion, fewer incoming messages, less variability in location patterns, and higher asymmetry of EDA between the right and left wrists.

References

[1] WHO, "World Health Organization news release," http://www.who.int/mediacentre/news/releases/2017/world-health-day, 1948, online; accessed April'17.

[2] S. N. Haber and S. L. Rauch, "Neurocircuitry: a window into the networks underlying neuropsychiatric disease," *Neuropsychopharmacology*, vol. 35, no. 1, p. 1, 2010.

[3] S. Kumar, G. D. Abowd, W. T. Abraham, M. al?Absi, J. G. Beck, D. H. Chau, T. Condie, D. E. Conroy, E. Ertin, D. Estrin *et al.*, "Center of excellence for mobile sensor data-to-knowledge (md2k)," *Journal of the American Medical Informatics Association*, vol. 22, no. 6, pp. 1137–1142, 2015.

[4] Empatica, "Empatica E4," https://store.empatica.com/products/e4-wristband, 2012, online; accessed May'17.

[5] N. Jaques, S. Taylor, A. Azaria, A. Ghandeharioun, A. Sano, and R. Picard, "Predicting students' happiness from physiology, phone, mobility, and behavioral data," in *ACII, 2015*. IEEE, 2015, pp. 222–228.

[6] M. Matthews, S. Abdullah, G. Gay, and T. Choudhury, "Tracking mental well-being: Balancing rich sensing and patient needs," *Computer*, vol. 47, no. 4, pp. 36–43, 2014.

[7] C. R. R. G. Telford C, McCarthy-Jones S, "Experience sampling methodology studies of depression: the state of the art," *Psychological medicine*, vol. 42, no. 6, pp. 1119–1129, 2012.

[8] D. C. Mohr, E. Montague, C. Stiles-Shields, S. M. Kaiser, C. Brenner, E. Carty-Fickes, H. Palac, and J. Duffecy, "Medlink: a mobile intervention to address failure points in the treatment of depression in general medicine," in *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*. ICST, 2015, pp. 100–107.

[9] M. Sung, C. Marci, and A. Pentland, "Wearable feedback systems for rehabilitation," *Journal of neuroengineering and rehabilitation*, vol. 2, no. 1, p. 17, 2005.

[10] G. Valenza, C. Gentili, A. Lanatà, and E. P. Scilingo, "Mood recognition in bipolar patients through the psyche platform: Preliminary evaluations and perspectives," *Artificial intelligence in medicine*, vol. 57, no. 1, pp. 49–58, 2013.

[11] O. Mayora, M. Frost, B. Arnrich, F. Gravenhorst, A. Grunerbl, A. Muaremi, V. Osmani, A. Puiatti, N. Reichwaldt, C. Scharnweber *et al.*, "Mobile health systems for bipolar disorder: the relevance of non-functional requirements in monarca project," in *E-Health and Telemedicine: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2016, pp. 1395–1405.

[12] S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, and D. C. Mohr, "Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study," *JMIR*, vol. 17, no. 7, p. e175, 2015.

[13] M. Rabbi, S. Ali, T. Choudhury, and E. Berke, "Passive and in-situ assessment of mental and physical well-being using mobile sensors," in *UbiComp 2011*. ACM, 2011, pp. 385–394.

[14] H. Dibeklioglu, Z. Hammal, and J. F. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," *IEEE Journal of Biomedical and Health Informatics*, 2017.

[15] S. Scherer, G. M. Lucas, J. Gratch, A. S. Rizzo, and L.-P. Morency, "Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 59–73, 2016.

[16] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 3–10.

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

[18] J. E. Broderick and G. Vikingstad, "Frequent assessment of negative symptoms does not induce depressed mood," *Journal of clinical psychology in medical settings*, vol. 15, no. 4, pp. 296–300, 2008.

[19] A. Guidi, S. Salvi, M. Ottaviano, C. Gentili, G. Bertschy, D. de Rossi, E. P. Scilingo, and N. Vanello, "Smartphone application for the analysis of prosodic features in running speech with a focus on bipolar disorders: system performance evaluation and case study," *Sensors*, vol. 15, no. 11, pp. 28 070–28 087, 2015.

[20] A. Maxhuni, A. Muñoz-Meléndez, V. Osmani, H. Perez, O. Mayora, and E. F. Morales, "Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients," *Pervasive and Mobile Computing*, vol. 31, pp. 50–66, 2016.

[21] MovisensXS, "eXperience Sampling for Android," https://xs.movisens.com, 2012, online; accessed April'17.

[22] H. F. S. Gamboa, "Multi-modal behavioral biometrics based on hci and electrophysiology," Ph.D. dissertation, Universidade Técnica de Lisboa, 2008.

[23] N. G. Ward, H. O. Doerr, and M. C. Storrie, "Skin conductance: A potentially sensitive test for depression," *Psychiatry Research*, vol. 10, no. 4, pp. 295–302, 1983.

[24] R. W. Picard, S. Fedor, and Y. Ayzenberg, "Multiple arousal theory and daily-life electrodermal activity asymmetry," *Emotion Review*, vol. 8, no. 1, pp. 62–75, 2016.

[25] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxeda: A convex optimization approach to electrodermal activity processing," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 797–804, 2016.

[26] M. Borazio, E. Berlin, N. Kücükyildiz, P. Scholl, and K. Van Laerhoven, "Towards benchmarked sleep detection with wrist-worn sensing units," in *Healthcare Informatics (ICHI), 2014 IEEE International Conference on*. IEEE, 2014, pp. 125–134.

[27] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988.

[28] J. Hernandez, M. E. Hoque, W. Drevo, and R. W. Picard, "Mood meter: counting smiles in the wild," in *UbiComp 2012*. ACM, 2012, pp. 301–310.

[29] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2014, pp. 3–14.