On Predicting Relapse in Schizophrenia using Mobile Sensing in a Randomized Control Trial

Rui Wang[1], Weichen Wang[1], Mikio Obuchi[1], Emily Scherer[1], Rachel Brian[2], Dror Ben-Zeev[2], Tanzeem Choudhury[3], John Kane[4], Martar Hauser[4], Megan Walsh[4], Andrew Campbell[1]
[1]*Dartmouth College,* Hanover, NH, USA; [2]*University of Washington,* Seattle, WA, USA;
[3]*Cornell University,* Ithaca, NY, USA; [4]*Northwell Health,* Glen Oaks, NY, USA
{rui.wang.gr, weichen.wang.gr, mikio.obuchi.gr, emily.a.scherer, andrew.t.campbell}@dartmouth.edu;
{dbenzeev, rbrian}@uw.edu; tanzeem.choudhury@cornell.edu; {jkane2, mhauster, mwalsh9}@northwell.edu

***Abstract*—Schizophrenia is a severe psychiatric disorder. We use the CrossCheck study dataset to develop methods to predict whether or not a patient with schizophrenia is going to relapse from mobile phone data. Out of 75 patients in the year long randomized controlled trial only 27 relapse episodes occur. We apply various techniques to address predicting rare events in a longitudinal dataset. We apply resampling methods combining oversampling relapse examples and undersampling non-relapse examples and impute missing data. To avoid overfitting, we apply feature selection and transformation (i.e., PCA) to reduce the feature dimensionality. We find the best relapse prediction result using the first 100 principal components from both passive sensing and self-reports with 30-day prediction windows (precision=26.8%, recall=28.4%). If we demand the recall to be greater than 50%, we find the best result using 25 principle components from both passive sensing and self-reports with 30-day prediction windows (precision=15.4%, recall=51.6%).**

***Index Terms*—mobile sensing, mental health**

I. Introduction

Serious mental illness, such as schizophrenia, schizoaffective disorder and severe forms of bipolar disorder typically involve psychosis, and as a result, require long-term clinical care and hospitalization. Psychosis impacts a person's ability to think clearly and deal with reality because of delusional behavior. As a result, the illness affects perception, cognition, emotion and behavior. Schizophrenia is a severe psychiatric disorder that develops in approximately 1% of the world's population [1]. Most people with schizophrenia move between periods of relative remission and episodes of symptom exacerbation, relapse and hospitalization. Evidence suggests that clinical intervention at an early enough stage is effective in the prevention of transitions into a full relapse state. This reduces the need for hospitalization and can also lead to faster returns to remission [2]. Existing clinical practices are inefficient in detecting early precursors of relapse. Standard methods are typically based on face to face interactions and assessments with clinicians, conducted at set times and locations.

In this paper, we advance the vision of data-driven psychiatry and predict relapse using mobile phone sensing data [3], [4] from a year-long study for patients with schizophrenia. Relapse is defined as one of the following events [5]: 1) psychiatric hospitalization, 2) increased frequency or intensity of services, 3) increased dosage / additional medication and 25% increase in BPRS (the brief psychiatric rating scale) [6] from baseline/last assessment, 4) suicidal ideation, 5) homicidal ideation, 6) self-injury, and 7) violent behavior resulting in damage to property or person.

We use the complete CrossCheck dataset [3] to predict relapse. The CrossCheck study is conducted at Zucker Hillside Hospital, New York City, a large psychiatric hospital, with 150 outpatients with schizophrenia for 12 months using rolling enrollment. The participants are randomized to one of two arms: CrossCheck smartphone arm (n=75) or treatment-as-usual arm (n=75). The CrossCheck app [3] continuously records participants' physical activities (e.g., stationary, in a vehicle, walking, etc.), sleep (duration, bed time, and rise time), and sociability (i.e., the number of independent conversations a participant is around and their duration). The app also collects audio amplitude, light sensor readings, location coordinates, app usages and call logs. The app uses a built-in EMA (ecological momentary assessment) component to administer self-reported EMAs. Relapses are determined by clinical assessors. In this paper, we focus on the relapse date and predict whether or not a participant is going to relapse the next day. Of the 75 participants enrolled in the CrossCheck smartphone arm, 61 completed data collection. Clinical assessors identify 27 relapse cases from 20 of the 61 patients that completed the study.

Previous work using the CrossCheck dataset identified statistically significant associations between sleep, mobility, conversations, phone usage features and self-reported indicators in schizophrenia. Using these features the authors [3] developed inference models capable of predicting aggregated EMA scores that relate to self-reported mental health indicators (e.g., seeing things, hearing voices, feeling depressed) with a mean error of 7.6% of the score range. In a follow up paper [4], the authors predict BPRS scores of patients using mobile sensing and self-reports. BPRS is a survey administered by clinicians to evaluate symptom severity in schizophrenia. In this paper, we advance the prior work [3], [4] by developing models to address the challenges of predicting whether or not a patient in the CrossCheck study is going to relapse from mobile sensing data. Most standard classification algorithms

assume a relatively balanced class distribution and equal misclassification costs. An imbalanced dataset violates such an assumption, which leads to poor classification performance. We apply various techniques to address these challenges. First, we apply resampling methods combining oversampling relapse examples and undersample non-relapse examples to the training dataset such that the number of relapse and non-relapse examples are the same. Second, we impute missing sensing data to make sure we have enough data to train the classifiers. Finally, we apply feature selection (e.g., L1 regularization) and feature transformation (i.e., PCA) to reduce the feature dimensionality.

To the best of our knowledge, we are the first to present results for predicting relapse in out-patients with schizophrenia using sensing data from mobile phones. The contribution of this paper is as follows:

- We study the efficacy of using passive sensing data and self-report EMAs to predict relapse. We present classification performance from using only EMA or sensing data, and a combination of EMA and sensing data.
- We discuss several data preprocessing techniques to overcome problems with a real-world relapse dataset including aggregating daily features in different prediction time windows, data cleaning, missing data imputation, feature space transformation and dimensionality reduction.
- We find the best relapse prediction result using the first 100 principal components (PCs) from both passive sensing and EMA with 30-day windows (precision=26.8%, recall=28.4%). Note, that a number of studies [7] find that most patients with schizophrenia experience symptoms 30 days before relapse. Our time window derived from sensing data confirms this known finding.

While the performance of our predictor shows how challenging the problem of relapse prediction is, a key contribution of the paper is what we discover in developing the relapse classifiers (e.g., per-participant standardization does not help with prediction, finding that a 30-day window offers the best prediction). Accurately predicting relapse is a difficult problem because relapses are rare. However, we find certain model design considerations help with the development of better models. Specifically, we find the prediction performance peaks with a 30-day prediction window. Transforming the features using PCA reduces the feature dimensionality and generates more useful features. The principal components reveal different behavioral patterns that are associated with relapses. Finally, self-reported EMAs are not good relapse predictors by themselves, but combining EMA and passive sensing data improves the performance.

II. Method

We aim to predict whether or not a participant relapses during the year-long span of the CrossCheck study using the smartphone passive sensing data and self-report EMAs. In what follows, we discuss the relapse dataset, data preprocessing, behavioral features computed from the passive sensing data, and prediction models in detail.

A. Dataset

The CrossCheck dataset is a rich, unique psychiatric dataset; it includes: mobile sensing data, pre-post surveys, weekly EMAs by the participants associated with their symptoms, BPRS scored clinically administered surveys across the year on a weekly to monthly basis depending on patients' condition severity, and the details associated with relapse of patients.

In the dataset, 61 out of 75 participants in the CrossCheck smartphone arm completed the full year-long study. One interesting insight is that while there were many issues of missing data, and lost and stolen phones during the study the vast majority of participants completed the study with good to high compliance. This strongly counters the occasionally aired opinion that people with serious mental illness can not adopt and use mobile technology. 26 of the 61 participants are female and 35 male. There are 24 African American, 5 Asian, 2 Multiracial, 29 Caucasian and 1 Unknown in the study. The average number of days a participant is in the study is 322 days (SD = 93, median = 361). We identify 27 relapses from 20 participants, in which 16 participant relapse once, 1 participant relapse twice, and 3 participants relapse three times each.

The CrossCheck app dataset includes a wide range of behavioral passive sensing data from the phone. Specifically, it includes physical activities, locations, ambient sound levels, voice/noise labels, number of calls and text messages, application usage, screen lock/unlock, and ambient light intensity. We compute features from the passive sensing data on a daily basis, which describe participant's behaviors (e.g., duration of different physical activities in a day, conversation duration and frequency, different types of places visited, app usage).

The dataset includes a 10-item EMA self-reported every Monday, Wednesday, and Friday. The EMA asks participants to score themselves on been feeling calm, social, bothered by voices, seeing things other people can't see, feeling stressed, worried about people trying to harm them, sleeping well, able to think clearly, depressed and hopeful about the future.

B. Behavioral Features

We incorporate passive sensing features proposed in [3], [4]. These features are predictive of self-reported and clinician-administered symptoms among schizophrenia patients. The features are computed on a daily basis and also broken down into four epochs of the day: *morning (6am-12pm)*, *afternoon (12pm-6pm)*, *evening (6pm-12am)* and *night (12am-6am)*. These epoch features allow us to model people's behaviors during different parts of the day (e.g., walking in the morning, sleeping in the afternoon, not socially engaged in the evening, using the phone a lot during the night period). Specifically, we compute the following features. To measure *physical activities*, we compute duration for different activities (e.g., on foot, still, in vehicle, and on bicycle), and in order to measure *mobility*, we compute the number of locations visited and distance traveled. To measure *sleep patterns*, we infer sleep duration, sleep start, and end time from sensing data, and to measure *ambient environmental context*, we compute the amplitude of

ambient sound and ambient light. We also compute *face-to-face conversations* features which consist of conversation frequency and duration, and *phone-usage* features including the number of phone calls, SMS, and lock/unlock frequency and duration. We also compute *semantic location* [8]. Specifically, we consider the following places: home, food, travel, art&entertainment, nightlife, education, parks&outdoors, library, shop, gym, medical and residence. We compute the time spent at these places every day. We first identify significant locations where a participant dwells for a significant amount of time of the day. We find significant locations by clustering the GPS coordinates collected in a day using density-based spatial clustering of applications with noise (DBSCAN) [9]. The centroid of each cluster is considered a significant location. We assume participants are usually at their homes sleeping between 2 am to 6 am. Therefore, we label a significant location as home where a participant spends most of the time between this period of the night. We then use the Foursquare API [10] to label the other significant locations.

C. Data preprocessing

In what follows, we describe our data preprocessing, which include aggregating daily features in different relapse prediction time windows, data cleaning, missing data imputation, feature space transformation and dimensionality reduction.

Relapse prediction time window. We define the relapse prediction window as the number of days before the day identified as the *start of a relapse*. Studies find that most patients with schizophrenia experience symptoms 30 days before relapse [7], [11], [12]. Therefore, we evaluate relapse prediction using four different time windows: 7 days, 14 days, 21 days, and 30 days. We summarize the daily features within the prediction window as the average value of each of the features. The prediction time window construction is illustrated in Fig. 1. Specifically, suppose the prediction window size is 7 days, we first identify the date of the first relapse, then we group the 7 days before the relapse day into a 7-day block and label the block as 1 (relapse). We compute the average value of every feature within the 7-day block. Then we group 7 days before the first day of the relapse block into a 7-day block and label the block as 0 (non-relapse). We repeat until the method reaches the beginning of the study. If the last block is shorter than 7 days, we discard the block. We discard 30 days of data after each relapse because many of the participants are hospitalized and can not have phones while on psychiatric units at the hospital. We repeat the above steps to group and label prediction windows for the rest of the data.

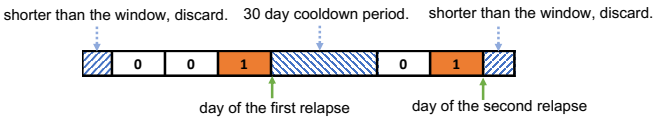


Fig. 1: Prediction window construction. Each window is labeled as 0 (non-relapse) or 1 (relapse in the following day).

Data cleaning and imputation. We compute behavioral features on a daily basis. Poor daily data quality may skew the behavioral features; we exclude the days with less than 19 hours of data. We use this threshold across this paper because we find lowering the threshold (e.g., missing 10 hours of data) does not include significantly more data whereas the data quality is poorer. Specifically, we compute the number of hours of data we have received for each passive sensing data. We label the sensing data as missing if less than 19 hours of data are collected in that day. We also control the data quality for aggregated time window features. We label the average feature values as missing if the feature misses over 70% of the days in the time window. We exclude time windows with more than 70% of feature values that are missing from our analysis. We heuristically pick the threshold to balance the data quality and make sure we have enough data for our analysis. The number of non-relapse and relapse cases are shown in Table I. We use a Singular Value Decomposition based method SVDimpute [13] to impute missing values. SVDimpute is a robust and sensitive method for missing value estimation surpassing the commonly used row average method.

Per-participant standardization. We use per-participant standardization to remove between-individual differences from the behavioral features. We hypothesize that different people may have different behavioral baselines. For example, a construction worker might be more physically active than an office worker whereas they have the same mental health outcomes (e.g., relapse). However, the within-individual differences in behaviors might be more inductive of changes in mental health. Per-participant standardization removes the between-individual behavioral differences and keeps only within-individual behavioral differences. We test our hypothesis in Section III.

Per-participant standardization transforms a participant's passive sensing features and EMA responses according to their first 30 days' data. Specifically, we first compute the mean μ_{30} and standard deviation σ_{30} for each of the features in the first 30 days, then we transform the feature as follows: $v_t = (v - \mu_{30})/\sigma_{30}$, where v is the original feature vector and v_t is the transformed feature vector. We apply per-participant standardization before aggregating features into prediction windows. We evaluate relapse prediction performance with or without per-participant standardization.

Feature space transformation and dimensionality reduction. We use principal components analysis (PCA) [14] to transform the feature space and reduce the feature dimensionality. PCA transforms a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The principal components are defined in a way that the first principal component accounts for as much of the variability in the data as possible, and each succeeding component in turn accounts for as much of the rest of the variability in the data as possible. The resulting principal components are an uncorrelated orthogonal basis set. The original observation (i.e., the feature values in a prediction

window) can be reconstructed by a linear combination of the principal components. We use the weights of each PC as transformed features. We can use a smaller number of PCs to reconstruct the original observation, which leads to a smaller number of features (i.e., reduce the dimensionality). *Each PC can be interpreted as a behavioral pattern*. For example, if a PC has large positive weight in the component for phone unlock duration and phone call duration features, and large negative weight for still duration, we would interpret this PC represents a high phone usage and high sedentary behavioral pattern.

We evaluate relapse prediction for different PCA setups. We first use the raw feature values to predict relapses. Then, we experiment with using different number of PCs to predict relapse. Specifically, we test using the first 1, 2, 5, 10, 25, 50, and 100 PCs, which explain 28.9%, 45.1%, 69.5%, 80.1%, 90.2%, 96.9%, and 99.9% of the variance in sensing and EMA data combined, to predict relapses.

D. Relapse Prediction as Binary Classification

Relapse prediction is a binary classification problem, i.e., we classify a n-day time window as non-relapse or relapse. We evaluate four popular classifiers: logistic regression, SVM with linear kernel [15], SVM with radial basis function kernel (RBF kernel) [16], and random forest [17]. The classifiers include linear classifiers (i.e., logistic regression, linear SVM), non-linear classifiers (i.e., RBF SVM, random forest), and non-parametric classifier (i.e., random forest). We apply elastic net regularization [18] on logistic regression and linear SVM to avoid over-fitting. The elastic net linearly combines the L1 and L2 penalties of the lasso and ridge methods. We use a grid search to find the best model hyper-parameters. We aim to find how different types of classifier perform in predicting relapses.

There are two major challenges in predicting relapses. First, we do not have a large amount of data to train a classification model. We have the most examples with 7-day prediction window, in which 1641 windows are non-relapses and 16 are relapses. Second and more importantly, the participants do not relapse frequently, therefore, our dataset is imbalanced. In fact, only 0.97% of the 7-day prediction windows are labeled as relapse. Training the classifiers without augmenting the dataset results in 0% of recall. Prediction in an imbalanced dataset requires a large dataset. A future work could collect a much larger dataset over many years to help build a reliable relapse predictor. In what follows, we discuss our method to address the challenges in detail.

Resample the training data. To reduce the data bias in the dataset (i.e., more non-relapses than relapses), we apply data resampling techniques to balance the dataset. Resampling techniques are widely used to address the bias in an imbalanced dataset (i.e., majority cases have higher weight than minority cases). We use Synthetic Minority Over-sampling Technique (SMOTE) [19] to balance the training set by over-sampling the minority class (i.e., relapse) and under-sampling the majority class (i.e., non-relapse). Instead of over-sampling the minority classes by replication, SMOTE creates "synthetic" minority examples. The synthetic minority examples are generated from k-nearest neighbors of the existing minority examples [19]. We use 5-nearest neighbors to generate synthetic minority examples. SMOTE has shown to be more effective than simple under-sampling and over-sampling methods.

We utilize 2-level 3-fold stratified cross validation (CV) to evaluate the relapse prediction performance. The top level CV evaluates the prediction performance and the second level CV selects model hyper-parameters. In order to avoid selecting a random seed that leads to impractically high or low prediction performance, we repeat the 2-level 3-fold stratified cross validation 5 times and report the average prediction performance metrics.

III. Results

In what follows, we discuss our relapse prediction results in detail. We first define our relapse prediction baseline, followed by the best results from each of the four classifiers (i.e., logistic regression, SVM with linear kernel, SVM with RBF kernel, and random forest). We then discuss how different classifier design considerations (i.e., using raw feature or standardized features, different data types, prediction window length and PCA) affect prediction performance. Finally, we present the features that are important to predict relapse.

A. Relapse prediction baseline

Because there is no prior work on using passive sensing to predict relapse in schizophrenia patients, we use random guessing as our prediction baseline. Specifically, we randomly label a case with either relapse or non-relapse with the same probability. Other simple prediction baselines (e.g., assign the same label to all examples) produce either 100% or 0% in recall, which is not informative than random guessing in this case because precision and recall are both important metrics in predicting relapse. We then compute precision, recall and F1 score for the random labeled cases. The baseline performance is presented in Table I.

TABLE I: Relapse prediction baseline according to random guessing for a classification.

window len	#non-relapse	#relapse	precision	recall	F1
7	1641	16	0.010	0.500	0.019
15	861	18	0.020	0.500	0.039
21	578	20	0.033	0.500	0.063
30	411	19	0.044	0.500	0.081

B. Results overview

Next, we present the best prediction results (i.e., highest F1 score) from the four classifiers obtained using grid search. Table II shows the corresponding prediction windows length, the number of principal components (PCs), precision, recall and F1 score that are associated with best performance from each of the classifiers. *Interestingly, all classifiers achieve the best F1 score using non-standardized data with 30-day*

prediction time window. We suspect behavioral patterns over a longer period (e.g., 30 days) are more indicative of future relapse. SVM with RBF kernel achieves the best F1 score among the four classifiers using the first 100 PCs obtained from both sensing and EMA data. The precision is 26.8% and the recall is 27.4%. To put these numbers into perspective, there are 411 cases in the 30-day dataset, 19 of which are relapses. The classifier predicts 19 cases that are relapses, 5 of which are correct and 14 are incorrectly identified as relapse. 14 relapses are misclassified as non-relapse. Logistic regression and SVM with linear kernel achieve slightly worse F1 scores but higher recall. The logistic regression model achieves 35.8% of recall and 21.4% of precision. The SVM with linear kernel achieves 32.6% of recall and 23.3% of precision. The random forest model achieves the worst F1 score, with 18.9% of recall and 28.1% of precision. All four classifiers beat the baseline in terms of the F1 score and precision. However, the recall is worse than the baseline.

TABLE II: Best prediction results according to the F1 score

data type	classifier	window len / # of PCs	precision/recall/F1
sensing+ema	svm rbf	30 / 100	0.268/ 0.284/ 0.274
sensing+ema	logistic regression	30 / 50	0.214/ 0.358/ 0.265
sensing+ema	svm linear	30 / 50	0.233/ 0.326/ 0.262
sensing	random forest	30 / 25	0.281/ 0.189/ 0.223

In summary, SVM with RBF kernel, SVM with linear kernel, and logistic regression achieve similar relapse prediction performance, whereas random forest achieves the worst performance. 30 days is the best time window to predict relapse. Combining passive sensing data and self-report EMA responses help predicting relapse. Standardizing every participant's data does not help improve performance. On the contrary, we observe poorer performance with standardized data. Using PCA to combine features and reduce the feature dimensionality improve performance. We discuss how using different data as predictors, prediction window, and PCA affects prediction performance in the following sections.

C. Prioritizing the recall

In the previous section, we present the best prediction results in term of the F1 score. The F1 score is the harmonic average of the precision and recall, which gives the same weight to precision and recall. However, misclassifying relapse as non-relapse may have severe consequences compared with misclassifying non-relapse as relapse. Misclassifying a relapse as non-relapse may lead to non-action (e.g., fail to deliver intervention) and miss the best opportunity to treat the patient, whereas misclassifying a non-relapse as relapse may lead to unnecessary intervention and clinical visits thus increased cost. These are important considerations in building a real-time relapse prediction and clinical intervention system as envisioned in [20]. While this paper represents a first dive into developing a predictor for relapse we believe that these performance figures while good considering the challenge of the problem would be significantly improved if we had more relapse examples. The study power analysis underperformed in the sample and the duration of the study - a larger N or longer study duration is part of future work. In what follows, we present prediction results with the constraint that recall $\geq 50\%$. The results are presents in Table III.

TABLE III: Best prediction results based on the F1 score with recall $\geq 50\%$

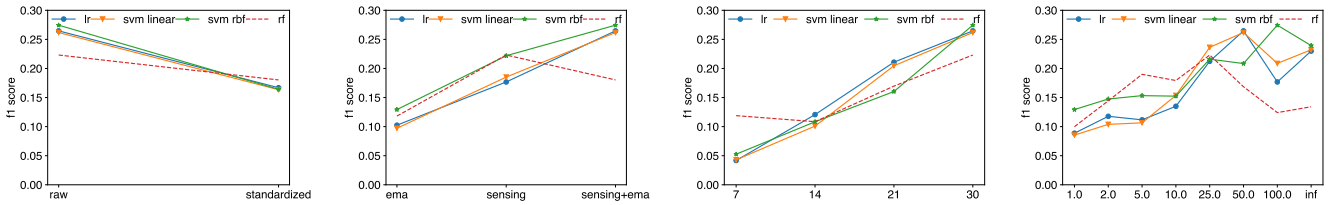
data type	classifier	window len / # of PCs	precision/recall/F1
sensing+ema	svm linear	30 / 25	0.154/ 0.516/ 0.236
sensing+ema	svm rbf	30 / 50	0.140/ 0.537/ 0.208
sensing+ema	logistic regression	30 / 2	0.068/ 0.505/ 0.118
ema	random forest	21 / 1	0.055/ 0.562/ 0.100

SVM with linear and RBF kernels, and logistic regression achieve the best F1 scores when recall $\geq 50\%$ using both sensing and EMA data with 30-day time window whereas random forest achieves the best result with 21-day window EMA data. All models beat the baseline in term of precision, recall and F1 score. The random forest model achieves the best F1 score using only EMA data as predictors with 21-day prediction window. However, the performance is only slightly better than the 21-day window baseline. We discuss the result from SVM with linear kernel in detail. The precision of the model is 15.4% and the recall is 51.6%. The classifier predicts 64 cases that are relapses, 10 of which are correct and 54 are incorrectly identified as relapse. 9 relapses are misclassified as non-relapse. Compared with the result with the best F1 score, the model correctly identifies more relapses with the cost of more false positives. The result shows that in practice, we can bias our models to be more sensitive to relapse with the cost of more false positives. To do so, we could adapt our model parameters by assigning different weights to precision and recall thus obtain a more desirable relapse prediction model.

D. Prediction performance analysis

In what follows, we discuss how different model decisions affect the relapse prediction performance. Specifically, we focus on whether or not we standardize each participant's features, what types of data are included in the prediction model (i.e., EMA, sensing, and both EMA and sensing), what prediction window we use (i.e., 7-day, 14-day, 21-day, or 30-day), and whether or not we apply PCA to transform the data and how many PCs we should use if PCA is applied.

Per-participant standardization transforms a participant's passive sensing features and EMA responses according to their first 30 days' data. We first compute the mean μ_{30} and standard deviation σ_{30} for each of the features in the first 30 days, then we transform the feature as follows: $v_t = (v - \mu_{30})/\sigma_{30}$, where v is the original feature vector and v_t is the transformed feature vector. Figure 2(a) shows the best F1 scores obtained from four classifiers with or without per-participant standardization. Applying per-participant standardization leads to poor F1 scores. Logistic regression and SVM models show



(a) standardization. (b) data types. (c) prediction window length. (d) # of PCs. inf: no PCA applied.

Fig. 2: Predicting F1 score from different models.

overall similar F1 scores, where the F1 scores decrease from 0.256 to 0.169 after applying per-participant standardization. The results show that per-participant standardization does not improve the performance. We suspect the absolute behavioral levels (e.g., sleep duration), which are eliminated by per-participant standardization, are helpful in predicting relapse.

Data types. We inspect how EMA responses and passive sensing data help predicting relapse. Figure 2(b) shows the best F1 scores obtained from four classifiers with three settings: 1) predict using only EMA responses, 2) predict using only passive sensing data, and 3) predict using both EMA responses. All classifiers perform poorly using only EMA responses, where the F1 scores are around 0.1. However, the F1 scores significantly improve when we use passive sensing data for prediction, where SVM with RBF kernel achieves the best performance with $F1 = 0.222, \text{precision} = 22.2\%, \text{recall} = 23.2\%$. Random forest achieves similar prediction performance. Logistic regression and SVM with linear kernel, however, perform poorer than RBF SVM and random forest. We suspect the non-linearity of the RBF kernel and random forest helps to reduce under-fitting. *We achieve the best prediction performance by combining both EMA and passive sensing.* SVM with RBF kernel achieves the best performance with $F1 = 0.274, \text{precision} = 26.8\%, \text{recall} = 28.4\%$. Logistic regression and linear SVM achieve slightly poorer performance whereas random forest achieves a poorer F1 score compared with using only sensing data. In summary, we predict relapse more accurately using passive sensing data compared with only using self-report EMA. Combining passive sensing data and EMA self-reports further improves the prediction performance.

Prediction window length. Figure 2(c) shows the best F1 scores obtained from four classifiers with four different prediction window settings: 7-day, 14-day, 21-day, and 30-day. The F1 scores increase for all classifiers as we increase the window length from 7 days to 30 days. We suspect behavioral patterns over a longer period are more indicative of future relapse. Therefore, we find better prediction results with longer prediction windows. However, increasing the prediction window length reduces the number of examples available for training and testing prediction models. Take the RBF SVM as an example, the classifier achieves $F1 = 0.053$ with 7-day window, which is 0.034 higher than the baseline shown in Table I, whereas it achieves $F1 = 0.274$ with 30-day window, which is 0.193 higher than the baseline. We suspect summarizing behavioral features in shorter windows leads to more noise in the feature data because of the short-term behavior changes whereas longer windows smooth the behavioral data so that the features captures participants' behaviors more accurately.

PCA. Figure 2(d) shows the best F1 scores obtained from four classifiers with different PCA settings. As we include more PCs in the predictions, the F1 scores increase for all classifiers. Two linear classifiers, linear SVM and logistic regression, achieve the best F1 scores when using 50 PCs, which is higher than using the raw feature data without PCA transformation. RBF SVM achieves the best F1 score when using 100 PCs, which again is higher than using the raw feature data. Random forest achieves the best F1 score when using 25 PCs, however, the prediction performance of random forest is poorer than the other three classifiers. The results show transforming the features using PCA reduces the feature dimensionality and generates more useful features by combining different features together. We discuss particular PCs in later sections.

E. Useful Features

In what follows, we present features selected by L1 regularization in logistic regression training. We do not transform the features using PCA so that we can interpret how features are related to relapse. We choose to present logistic regression coefficients instead of other classifiers because it is easier to interpret parameters in logistic regression - positive coefficients indicate positive correlations whereas negative coefficients indicate negative correlations. We first present features selected by the model using both sensing and EMA data, then, we present features selected by the model using only sensing data. The selected features and their regression coefficients are presented in Table IV.

Sensing and EMA. The logistic regression model achieves 22.6% precision and 36.8% of recall using both sensing and EMA data. The prediction window is 30 days. The L1 regularization selects 81 out of 144 features in training. We present the top 10 features with the largest absolute coefficients. We find that participants who have more conversations in the morning, walk more in the evening but visit fewer places in the evening, visit fewer educational, travel, and residential places, self-report a lower score in seeing things, but spend more time responding to EMAs are more likely to relapse.

TABLE IV: L1 selected features in logistic regression.

sensing+EMA (precision=22.6%, recall=36.8%)	sensing (precision=12.5%, recall=42.1%)
feature (coeff)	feature (coeff)
convo duration morning (2.631)	convo duration morning (0.659)
on foot duration evening (2.553)	voice ratio night (-0.319)
voice ratio morning (2.540)	number of calls made (-0.251)
visit education places (-2.139)	on foot duration evening (0.193)
# of visited places evening (-1.952)	visit outdoors places (0.084)
EMA response time (1.876)	call duration morning (-0.008)
EMA (seeing things) (-1.650)	
audio amplitude afternoon (-1.620)	
visit travel places (-1.597)	
visit residence places (-1.506)	

Please note, the L1 regularization also selects 5 EMA items (i.e., depressed, calm, voices, think, harm) and positive scores to predict relapse. Specifically, participants who self-report higher scores in depressed, voices, and harm items, and lower scores in calm, and thinking are more likely to relapse.

Sensing only. The logistic regression model achieves 12.5% of precision and 42.1% of recall using sensing data only. The prediction window is 30 days. The L1 regularization selects 6 out of 130 features in training. Specifically, participants who have more conversations in the morning, spend more time walking in the evening, visit more parks and outdoor places, makes fewer phone calls are more likely to relapse.

F. Behavioral Principal Components

In what follows, we present the top 5 PCs with the largest absolute logistic regression coefficients as shown in Table V. The top 5 PCs, their regression coefficients, and characteristics. A positive coefficient indicates the PC is positively correlated with relapse (i.e., larger PC weight indicates a higher probability of relapse). and a negative coefficient indicates a PC is negatively correlated with relapse.

PC 1 describes a behavioral pattern in which participants spend less time responding to EMAs, report lower scores in all EMA items, and makes more phone calls. *Participants whose behaviors are similar to PC 1 are less likely to relapse.*

PC 19 describes a behavioral pattern in which participants visit more places, spend less time at nightlife, arts and entertainment, parks and outdoor, and gym places, spend more time at residence, medical and education places, receive more phone calls but do not make many calls, have more conversation during the evening, and spend more time responding to EMA questions. *Participants whose behaviors are similar to PC 19 are more likely to relapse.*

PC 7 describes a behavioral pattern in which participants make and receive more calls, have less conversation in the morning but visit more places in the evening. *Participants whose behaviors are similar to PC 7 are less likely to relapse.*

PC 36 describe a behavioral pattern in which participants visit fewer places related to medical, gym, library but visit more places relate to arts and entertainment, home, residence. They report high scores in hearing voices, harm, and feel less hopeful. *Participants whose behaviors are similar to PC 36 are more likely to relapse.*

TABLE V: PCs with the largest absolute coefficients.

PC	coefficient	features
1	-437.7	low EMA item scores, respond EMAs fast, more phone calls especially in the evening and night, ride bikes
19	395.8	visit more places; visit fewer places relate to nightlife, arts and entertainment, parks and outdoor, and gym; visit fewer places relate to residence, medical and education; respond EMAs slow; receive more phone calls but make fewer phone calls in the afternoon; have more conversations in the evening
7	-348.0	make and receive more calls, have less conversation in the morning, visit more places in the evening, ambient light is bright at night.
36	307.0	visit fewer places relate to medical, gym, and library; visit more places relate to arts and entertainment, home, and residence; report higher EMA score in items including hearing voices, harm, and feel less hopeful
8	305.7	visit more places relate to medical; more SMS use at night; more conversation at night; fewer phone calls in the morning; bright at night; more phone use at night; wake up late.

PC 8 describe a behavioral pattern in which participants visit more places related to medical assistance, use SMS more, phone use through the day, and conversations at night, fewer calls in the morning, bright light conditions during at night, and wake up late. *Participants whose behaviors are similar to PC 8 are more likely to relapse.*

IV. Discussion

In what follows, we discuss our results. Our results show that per-participant standardization fails to improve prediction performance. The standardized feature values indicate how many standard deviations the true value is from the feature mean value, which measures the within-individual behavior differences. Applying per-participant standardization for relapse prediction assumes similar deviations from a participant's average behaviors across all participants account for relapse. However, our results show that applying per-participant standardization leads to poorer prediction performance, *which may indicate that the between-individual differences are more predictive of relapse than within-individual differences.*

We find the 30-day time window is the best time window to predict relapse. Summarizing behavioral features in shorter windows leads to more noise in the feature data because of the short-term behavior changes, whereas longer windows smooth the behavioral data. Also, the pre-relapse behavioral changes might be gradual and the behavioral changes may start many days before the relapse. A shorter time window may lead to lower resolution of changes in behaviors (i.e., behaviors in consecutive time windows are similar despite one is relapse the other one is not), thus more challenging for the classifiers to identify relapse signals. Note, that a number of studies [7] find that most patients with schizophrenia experience symptoms 30

days before relapse. Our time window derived from sensing data confirms this known finding.

We show that using passive sensing data greatly improves the prediction performance compared with using only self-report EMA. Combining both passive sensing and EMA further improves the prediction performance. *Our results indicate that passive sensing data has the potential to unobtrusively monitor and predict relapse in the future.*

Semantic location features, phone calls and conversational features are good predictors of relapse. Transforming the behavioral feature space using PCA provides more "interpretable" behavioral patterns. Our findings show that behavioral features that are more interpretable are more likely to be indicative of relapse. Future work should focus on designing features that are interpretable and capture people's higher level behaviors (e.g., go to work, socializing with friends, exercising) by combining different sensor streams. Exploring behavioral patterns (e.g., behavioral principle components) would further give more insight into relapse episodes. *It is very important that the results from mobile sensing and machine learning are interpretable by clinicians working in the field.*

Misclassifying relapse as non-relapse may have severe consequences compared with misclassifying non-relapse as relapse. Misclassifying a relapse as non-relapse leads to non-action (e.g., fail to deliver the intervention) and miss the best timing to treat the patient, whereas misclassifying a non-relapse as relapse may lead to unnecessary clinical visits thus increased cost. We show that our models achieve 53.7% of recall with the cost of lower precision, which is still better than the baseline defined in Section III-A. *For practical use, we need to carefully evaluate the precision recall trade off and select the best model that maximize recall (i.e., identify as many patients at risk as possible) while minimizing the unnecessary cost due to misclassifying non-relapses as relapses.*

The relapse prediction model presented in this paper shows great promise in using passive sensing to predict relapse. The models show reasonable performance using passive sensing and self-reports as well as just using passive sensing. A future real-time relapse system based purely on passive sensing opens the way for continuous assessment of relapse.

V. Conclusion

To the best of our knowledge, this is the first paper to present results for predicting relapse in outpatients with schizophrenia using passive sensing data from mobile phones. We presented and evaluated different prediction model design considerations and found that linear models (e.g., logistic regression and linear SVM) using PCA-transformed passive sensing and self-report EMA features best predict relapse with 30-day time window. We discussed the features and behavioral patterns that are predictive of relapse. Although our prediction performance present challenges to be deployed today in clinical practices, our results show promises in using passive sensing to help clinicians better identify patients at risks of relapse. We hope this paper leads to better designed studies, datasets and predictive models for relapse and ultimately helps transition psychiatry to be more data-driven.

References

[1] T. Vos, R. M. Barber, B. Bell, A. Bertozzi-Villa, S. Biryukov, I. Bolliger, F. Charlson, A. Davis, L. Degenhardt, D. Dicker *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the global burden of disease study 2013," *The Lancet*, vol. 386, no. 9995, pp. 743–800, 2015.

[2] R. Morriss, I. Vinjamuri, M. A. Faizal, C. A. Bolton, and J. P. McCarthy, "Training to recognize the early signs of recurrence in schizophrenia," *Schizophrenia bulletin*, vol. 39, no. 2, pp. 255–256, 2013.

[3] R. Wang, M. S. H. Aung, S. Abdullah, R. Brian, A. T. Campbell *et al.*, "Crosscheck: Toward passive sensing and detection of mental health changes in people with schizophrenia," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '16. ACM, 2016, pp. 886–897.

[4] R. Wang, W. Wang, M. S. Aung, D. Ben-Zeev, R. Brian, A. T. Campbell, T. Choudhury, M. Hauser, J. Kane, E. A. Scherer *et al.*, "Predicting symptom trajectories of schizophrenia using mobile sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, p. 110, 2017.

[5] J. G. Csernansky, R. Mahmoud, and R. Brenner, "A comparison of risperidone and haloperidol for the prevention of relapse in patients with schizophrenia," *New England Journal of Medicine*, vol. 346, no. 1, pp. 16–22, 2002.

[6] J. E. Overall and D. R. Gorham, "The brief psychiatric rating scale," *Psychological reports*, vol. 10, no. 3, pp. 799–812, 1962.

[7] M. Birchwood, J. Smith, F. Macmillan, B. Hogg, R. Prasad, C. Harvey, and S. Bering, "Predicting relapse in schizophrenia: the development and implementation of an early signs monitoring system using patients and families as observers, a preliminary investigation," *Psychological Medicine*, vol. 19, no. 03, pp. 649–656, 1989.

[8] W. Wang, S. Mirjafari, G. Harari, D. Ben-Zeev *et al.*, "Social sensing: Assessing social functioning of patients living with schizophrenia using mobile phone sensing," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 2020.

[9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *KDD '96*. AAAI Press, 1996, pp. 226–231.

[10] "Foursquare," https://developer.foursquare.com/places-api, 2018.

[11] Y. Henmi, "Prodromal symptoms of relapse in schizophrenic outpatients: retrospective and prospective study," *Psychiatry and clinical Neurosciences*, vol. 47, no. 4, pp. 753–775, 1993.

[12] N. Tarrier, C. Barrowclough, and J. Bamrah, "Prodromal signs of relapse in schizophrenia," *Social Psychiatry and Psychiatric Epidemiology*, vol. 26, no. 4, pp. 157–161, 1991.

[13] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[14] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *International Conference on Artificial Neural Networks*. Springer, 1997, pp. 583–588.

[15] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[16] Y.-W. Chang, C.-J. Hsieh, K.-W. Chang, M. Ringgaard, and C.-J. Lin, "Training and testing low-degree polynomial data mappings via linear svm," *Journal of Machine Learning Research*, vol. 11, no. Apr, pp. 1471–1490, 2010.

[17] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[20] H. Hsin, M. Fromer, B. Peterson, C. Walter, M. Fleck, A. Campbell, P. Varghese, and R. Califf, "Transforming psychiatry into data-driven medicine with digital measurement tools," *npj Digital Medicine*, vol. 1, no. 1, p. 37, Aug. 2018. [Online]. Available: https://www.nature.com/articles/s41746-018-0046-0