DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks

Yoshihiko Suhara[*] [†]
Recruit Institute of Technology
444 Castro Street Suite 900
Mountain View, CA 94041
suharay@recruit.ai

Yinzhan Xu [*]
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
xyzhan@mit.edu

Alex 'Sandy' Pentland
MIT Media Lab
20 Ames Street
Cambridge, MA 02139
pentland@mit.edu

ABSTRACT

Depression is a prevailing issue and is an increasing problem in many people's lives. Without observable diagnostic criteria, the signs of depression may go unnoticed, resulting in high demand for detecting depression in advance automatically. This paper tackles the challenging problem of forecasting severely depressed moods based on self-reported histories. Despite the large amount of research on understanding individual moods including depression, anxiety, and stress based on behavioral logs collected by pervasive computing devices such as smartphones, forecasting depressed moods is still an open question. This paper develops a recurrent neural network algorithm that incorporates categorical embedding layers for forecasting depression. We collected large-scale records from 2,382 self-declared depressed people to conduct the experiment. Experimental results show that our method forecast the severely depressed mood of a user based on self-reported histories, with higher accuracy than SVM. The results also showed that the long-term historical information of a user improves the accuracy of forecasting depressed mood.

Keywords

Depression; Neural Networks; Mobile Applications

1. INTRODUCTION

Depression is a prevailing mental health care problem and is a popular keyword due to the increase in mentally disordered patients including potential numbers. The WHO estimates that 676 million people in the world (nearly one in ten people) suffer from depression [1]. Current predictions by the WHO indicate that by 2030 depression will be the leading cause of disease burden globally. In the United States, mental disorder problems are among the top five conditions for direct medical expenditure, with associated annual health care costs exceeding $30 billion [29].

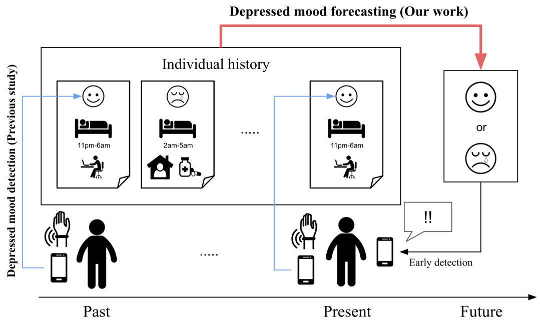


Figure 1: Depressed mood forecasting task.

Due to the lack of physical symptoms, diagnosing depression is a challenge. People cough or may have a fever when they are physically ill; these symptoms can lead them to go to hospitals for appropriate treatment. Without these physical symptoms, the signs of depression may go unnoticed. Existing studies [24][35] show effective vital signs for depression detection. A standard method is to measure biomarkers such as serotonin to provide obvious evidence for depression. However, obtaining such biomarkers requires special apparatus and often invasive sensing. Not many people can adopt the approach to assess their mental health in daily life. Because we lack a system to reveal mental illness through physical signs, we need an external system that helps us detect depression in a noninvasive manner. This need has led to a large body of work on depressed mood detection by pervasive computing devices [3][5][10][30][36][48][50].

Early detection of depressed mood is essential to provide appropriate interventions for preventing critical situations. Despite a large amount of research on existing depressive mood prediction [3][5][10][30][36][48][50], forecasting depressed moods has not been well studied. Therefore, we define a novel task of *forecasting depressed mood* and develop a predictive model for this task. In this paper, we distinguish forecasting from prediction, emphasizing the meaning of predicting future mood instead of existing mood. Particularly, we focus on forecasting *severe depression* among several types of depressive moods in this paper. Severely

[*] These authors contributed equally to this work.

[†] The author is also affiliated with MIT Media Lab.

[1] `http://www.who.int/healthinfo/`

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052676

depressed moods might cause irreversible damage or suicide attempts among individuals; thus, we consider the early detection of severe depression as essential to health care.

Figure 1 shows the methodological relationship between this paper and previous studies. Previous studies have solved the problem of predicting mood based on individual behavioral information collected by pervasive computing devices such as smartphones. We aim to forecast severely depressive mood based on individual histories including existing mood, behavioral information, and sleeping hours. To advance previous studies, this paper assumes that the information used to forecast depressive mood can be estimated accurately. Thus, we use self-reported histories instead of estimated information. Our focus is to understand the relationship between individual histories, including mood information in and up to the present, and a person's mood in the future.

Our first research question is "Can we forecast severe depression based on individual histories" (RQ1). The question includes what kind of histories contribute to improving forecasting performance. The second research question is "How many days do we need to look back to forecast severe depression?" (RQ2). The question is derived from two motivations. The first motivation is to confirm whether incorporating distant histories improves forecasting accuracy. If individual histories in a shorter period perform well, we do not have to use long-past histories in the forecast. The second motivation is to reveal the relationship between individual mood at some point and person's moods in the past; in other words, individual mental status is not always expressed by simple Markov state models. Instead, mental status should depend on complicated patterns of an individual's history. People usually have an experience in which a "subtle" event in the past strongly affected their mood at some time. To the best of our knowledge, no empirical study has answered RQ2.

In this paper, we developed a smartphone application for self-declared depressed people to collect moods, behavioral types, and medication records in a self-reported manner. We published the smartphone application in Google Play[2] and the App Store[3] in November 2012. By September 2014, the application had been used by 24,211 users. For our depression forecasting study, we filtered out users who had not used the application more than 28 successive days to ensure adequate history duration for each user. After this preprocessing step, the total number of users in the dataset[4] is 2,382 and the total number of recorded days is 345,158. To the best of our knowledge, this is the largest dataset that researchers have used for a mental health study.

This paper describes a deep learning algorithm we developed based on long short-term memory recurrent neural networks (LSTM-RNNs) [14] for using individual histories as time series features. LSTM-RNN is known for its capability for long-distance dependencies and is considered as a standard algorithm for a prediction task based on time series features. Our method separately prepares embedding layers for each categorical variables including a day-of-the-week variable in order to introduce day-of-the-week effect into the method.

Our contributions in the paper are as follows:

- We tackle the severe depression forecasting problem with a large-scale longitudinal dataset collected from 2,382 users over 22 months to take an essential step toward automatic and early depression detection.
- We develop an RNN-based algorithm for the depressed mood forecasting task. The algorithm introduces a day-of-the-week variable directly in order to take the day-of-the-week effect into account.
- The experimental results show that individual histories from the previous two weeks are indicative to forecast severe depression.

2. RELATED WORK

Automatic mood detection has been well studied. A major stream of automatic mood detection is a pervasive computing approach. Previous studies have used individual physical activity information including behavioral, mobility, sleeping [38], voice acoustic [25][34], and social patterns [28][26][32] collected by pervasive computing devices to estimate a person's mood, including depression [3][5][30][10][36][48][50], stress [3][4][12][39], anxiety [7][15], or well-being [16][18][21][33]. These studies use mental health assessment questionnaires such as the PHQ-9 [10] or EMA [18] to prepare ground truth information for building predictive models based on machine learning techniques.

Mehrota et al. [30] used phone usage patterns such as calls, SMSs, and overall application usage logs to calculate the correlation between the social interaction factors and depressive states. Farthen et al. [9] used the StudentLife dataset to build a depression classification that predicts users' PHQ-9 score based on features extracted from behavioral logs collected through smartphones. They categorized the users in the dataset into three groups and formalized the prediction task as a multiclass classification problem. Canzian and Musolesi [6] studies depression prediction only based on individual mobility traces. They analyzed the correlation between mobility metrics in the past time and PHQ-8 [41] scores of participants. They used SVM [8] to build a predictive model for predicting depression. MoodScope [22] collected a user's moods, communication, and phone usage logs to predict his or her mood based on a regression model. It used a dataset collected from 32 participants over two months in order to build a mood prediction model based on the collected information. Purple Robot [36] was developed to collect a user's location, movement, and phone usage in addition to responses to a self-reported depression survey. The authors conducted an experiment to collect a dataset from 40 participants for four months. They applied a linear regression model to show correlated features to understand the informative indicators of participants' depressive moods.

It is commonly known that individual depressive moods vary according to the day of the week. However, not many studies have empirically shown evidence of this. According to visitors to DepressedTest.com[5], which provides an online depression test, Monday is the most depressing day and Saturday is the least depressing day of the week. MacKerron and Mourato [27] conducted an experiment using a smartphone application called Mappiness[6] to collect 100,000 data

[2] https://play.google.com/store/apps/details?id=com.jp.hitrecord

[3] https://itunes.apple.com/us/app/utsureko-utsu-bingno-rennotameno/id576634652

[4] The users agreed to the data usage policy on the application before they started using the application.

[5] http://www.depressedtest.com/

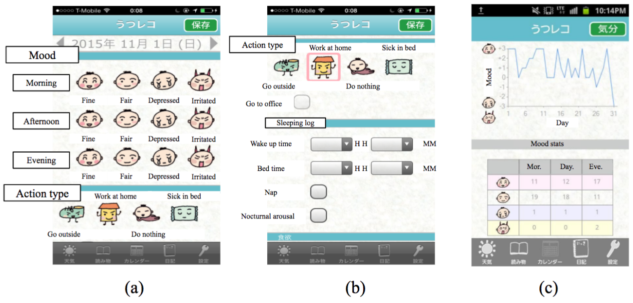
[6] http://www.mappiness.org.uk/

points of self-reported happiness. The results show that Tuesday is the least happy day of the week, followed by Monday. However, these studies do not take into account different times of the day. Golder and Macy [13] used millions of Tweets to estimate the trend of individual hourly moods over a week. They confirmed that Tuesday is the least happy day and Saturday morning is the least negative time period. However, because no study has empirically shown the trend of individual moods based on a large-scale dataset, we analyzed the trend of individual moods at three times of the day (morning, afternoon, and evening) to confirm the trend empirically. We also directly incorporated day-of-the-week information as a feature of our predictive model.

Several longitudinal experiments have been conducted for projects that aim to understand the relationship between human behaviors and mental health. The Friends and Family Study [1] is the first work in understanding how social dynamics affect many aspects of lives including health and moods. The study used an Android application called Funf [7] to collect call logs and SMS for communication, GPS trajectories for mobility, Bluetooth proximity for face-to-face communication, and smartphone usage. Weekly self-reported surveys were conducted to collect individual health statuses and moods. The 130 participants in the study were either couples or families who were members of a young-family residential living community adjacent to a major research university in North America. A part of the study, Moturu et. al. [32] showed that social interaction revealed their moods. The StudentLife Study [49] conducted a study that uses passive and automatic data collected from a class of 48 Dartmouth students via their smartphones over 10 weeks to assess their mental health. The study collected stress levels and positive affect information from participants to understand fine-grained mood transitions throughout the semester. The study revealed how academic activity affected students' affective moods [49]. The SNAPSHOT Study [8] collected sleep, network, affect, performance, stress, and health information from 200 socially connected undergraduate students over 30 days. It showed that predicting multiple factors such as depressed mood, happiness, health, and anxiety simultaneously via multitask learning improved prediction accuracy [17], indicating these factors share unified indicators. It also showed that sleep irregularity is strongly correlated to individual moods [37].

These existing studies tried to *detect* existing moods, including depression, based on social interaction, behavioral, and sleeping patterns. To the best of our knowledge, no study has tried to apply machine learning to build a predictive model to forecast depressive moods based on individual histories.

Our paper differs from existing studies in four ways: (1) This is the first work to build a predictive model to forecast individuals' depressed moods based on individual histories. (2) We use a large-scale dataset collected from more than 2,000 depressed users without limiting the participants' diversity. The dataset should cover a wide variety of depressed people in order to conduct a general depression study. (3) This is the first mental health study to apply a state-of-the-art deep learning technique to a prediction problem. It enables us to take a step forward in understanding how the features used for predictive models contribute to forecasting depressive moods. (4) We directly incorporate the day-of-the-week effect to improve the performance of the predictive model.



	Mor.	Day.	Eve.
	11	12	17
	19	18	11
	1	1	1
	0	0	2

Figure 2: Screenshots of the application used for collecting self-reported histories. (a) The application provides users with a single-tap interface to report moods and action types throughout the day. (b) Users can also submit sleeping times (bedtime, wake-up time) in a similar manner. (c) The application visualizes a user's history to look back the user's mood transition and maintain his or her motivation for using the application.

3. DATA COLLECTION

We developed a smartphone application called *Utsureko* for collecting data from users. Figure 2 presents screenshots of the application. The application provides an intuitive interface for users to input their moods in different time slots (morning, afternoon, and evening) each day. It also allows users to record their action type (i.e., go to office, go to work, work at home, do nothing at home, and sick in bed) and sleeping time, including bedtime and wake-up time. Users can input and update their records voluntarily at any time. The recorded information can be visualized in an aggregated manner so that users can look back at their records to capture general trends.

The application design followed existing psychiatric studies. Many mental assessment tools are described in the clinical psychology literature [20][43][41][40]. A common method is to ask a person to recognize his or her feelings via predefined questionnaires such as the PHQ-9 [40] and GAD-7 [40]. One example question from the PHQ-9 is as follows: "Over the last two weeks, how often have you been bothered by any of the following problems?—Little interest or pleasure in doing things (0, 1, 2, 3)." People answer the questions with multigraded scores. This approach ensures that people answer questions in a consistent manner. However, the major disadvantage of this approach is that recalled information is influenced by reconstructive processes that reduce the information's accuracy [31]. A recent trend, therefore, is to adopt an ecological momentary assessment (EMA) approach [43] to conduct mental assessments. The concept of EMA is to use open-ended questions to capture an individual's mental status close to the time that symptoms are detected. The advantage of EMA is that it tracks users' mental statuses with more fine-grained resolution than conventional mental health assessment tools such as PHQ-9. However, it has dis-

[7] http://funf.org/

[8] https://snapshot.media.mit.edu/

Table 1: Log information.

Category	Name	Value type
Mood	Morning mood	{fine, fair, depressed, irritated}
	Afternoon mood	
	Evening mood	
Behavioral log	Action type	{Go to work, go outside, work at home, do nothing at home, sick in bed}
	Medication	{yes, no}
	Urgent medication	{yes, no}
	Hospital attendance	{yes, no}
Sleeping log	Bedtime	HH:MM
	Wake-up time	HH:MM
	Nocturnal awakening	{yes, no}
	Taking a nap	{yes, no}

advantages for longitudinal participation, including that it is time-consuming to ask a user to input feelings every time he or she feels depressed.

We investigated the trade-off and concluded that a hybrid of conventional assessment and EMA was a good solution to develop our smartphone application to collect individual self-reported histories. The application has three aims: (1) to ask users predefined closed questions with single/multigraded answering options; (2) to collect individual moods at three periods of a day; (3) and to let users record their histories whenever they want. We decided to adopt aim (3) to collect longitudinal logs from a large number of depressed people without burdening them. In this paper, we also analyze the distribution of submission time to confirm that users periodically record their moods at different periods of a day.

Table 1 describes the collected self-reported information. The first category reflects the user's moods. We designed four categories of moods (fine, fair, depressed, and irritated). In addition to the depressed option, we prepared an irritated option, because uncontrollable irritability and anger are caused by certain types of depression [2]. Moods can be separately reported for three different parts of a day—morning, afternoon, and evening. Collecting moods at different periods of the day aims to capture the fine-grained transition of moods of users.

4. METHOD

4.1 Preliminaries

Definition 1. A user experiences a *severe depression day* if the user has negative feelings all day and exhibits inactive behavior in which he or she avoids leaving home.

Definition 2. (n, k)*-day severe depression forecast* is a forecasting task that classifies whether a user will experience at least one severe depression day in the coming n days based on the user's history during the last k days.

Problem Formulation. Given individual histories, the task is to forecast the existence of severe depression day in n days. For instance, (1, 14)-day severe depression forecast is used to forecast a user's severe depression on a coming day based on the histories of the last two weeks.

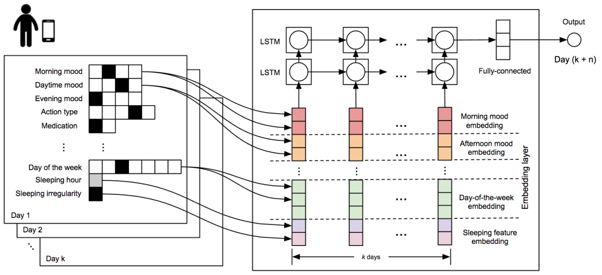


Figure 3: Network architecture of our method.

User histories are considered to be the time series of user logs. A feature vector extracted from the input of a user u in day t is denoted as $\mathbf{x}_t^u \in X^u$, where X^u denotes a sequence of feature vectors of user u. The severe depression label y_t^u is derived from user logs of day t if $n = 1$, or otherwise a period from t to $t + n - 1$.

Severe depression forecasting aims to forecast the severe depression of an individual user y_t^u based on the user's histories in the previous k days $(\mathbf{x}_{t-k-1}^u, \mathbf{x}_{t-k}^u, \ldots, \mathbf{x}_{t-1}^u)$. For instance, $k = 14$ is set to use information from the last two weeks for depression forecasting. Commonly used mental health questionnaires such as the PHQ-9 and GAD-7 asks patients how they have felt over the last two weeks to answer the items. Thus, in this paper, we set $k = 14$ and evaluate the forecasting performance by changing k values to answer RQ2.

4.2 Our method

With our large dataset, we used a supervised machine learning technique to build a predictive model for forecasting severe depression. To use individual histories as time series data, we needed a technique that was capable of incorporating dependencies from previous states. Among the techniques that are capable of handling sequential features, RNNs with hidden LSTM units [14] are known to be powerful models for learning from sequential data. They effectively model varying length sequences and capture long-range dependencies. Following the application of RNNs to clinical diagnoses classification [23], we present the first study to evaluate the ability of LSTM to forecast severe depression. The method also recognizes patterns in time series of self-reported user histories by visualizing representative patterns of nodes in a hidden layer.

The network architecture of our method is described in Figure 3. The architecture passes over all inputs in chronological order from $t - k - 1$ to $t - 1$ and generates an output $\hat{y}_t$ at the final sequence step of t. The LSTM layer is used to propagate historical information along with the features of the next step until it reaches the fully connected layer. The architecture can have more than one LSTM layer to empower the capability of long distance dependencies. The fully connected layer receives the propagated information. We apply the dropout technique [42] to avoid an overfitting problem. Because we have multiple categorical variables with distinct semantics, we introduce embedding layers for each categorical value to the model in order to convert categorical variables into a dense vectorial representation. This not only improves the model's performance but also provides a semantic interpretation of categorical variables

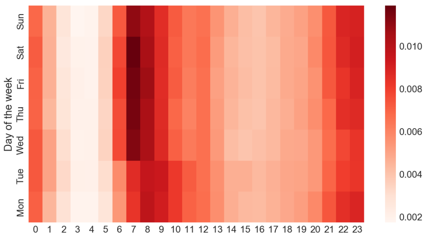


Figure 4: Distribution of user input hours for each day of the week.

[47]. Our model prepares the same number of nodes as the number of categories in each embedding layer. Features are extracted from collected logs shown in Table 1. We extract bedtime and wake-up time into sleeping hour and sleeping irregularity features. Sleeping hour is divided into 24 bins. We create a sleeping irregularity feature to capture the irregularity in sleeping hours, following previous studies that showed the relationship between sleeping irregularity and depressive moods. The sleeping irregularity feature takes 1 if the sleeping hour during day t is more different than 3 hours compared to the sleeping hour during day $t-1$, and 0 otherwise.

In addition to the categorical variables appearing in individual histories, we introduce a day-of-the-week variable and add a corresponding embedding layer to the model. Conventional studies have shown that people's mental health depends on a day of the week [13][44]. Thus, incorporating day-of-the-week information improves the model performance. Furthermore, it enables us to distinguish the same mood when it is experienced during the same period of time on different days of the week. For instance, being depressed on Friday evening should have a different meaning compared to being depressed on Sunday evening.

5. EVALUATION

In this section, we analyze the collected dataset to confirm that the dataset is reliable to conduct our study. Then we describe the experimental results and discuss our two research questions.

5.1 Dataset Analysis

Figure 4 presents the distribution of user input hours in the 7 day x 24 hours matrix. The matrix shows that users input data into their logs in the morning (6 a.m.–12 p.m.) and at night (9 p.m.–12 a.m.), indicating that users record their logs accordingly. This ensures that our application design is successful in keeping users' motivation to log their histories voluntarily.

Figure 5 shows the distribution of moods at each part of the day. Higher values denote more positive feelings during the periods. We confirm that the general trend of "afternoon > evening > morning" and "weekend > weekday," except for Sunday evening. The drop in the positive feeling ratio on Sunday evening shows so-called *Sunday night blues*. This coincides with the results estimated from millions of Tweets in [13]. The figure also supports the necessity of collecting user records for each part of the day separately because individual mood strongly depends on a part of the day.

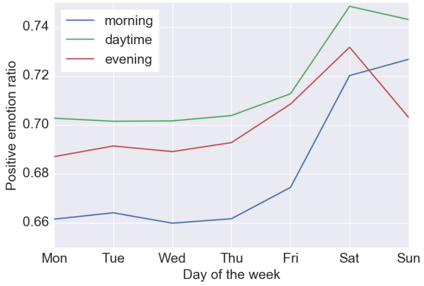


Figure 5: Positive feeling ratio for each day of the week. Each line corresponds to a time period in a day. The positive feeling ratio denotes the ratio of the count of positive feelings (fine or fair) to the total count of any feeling by all users.

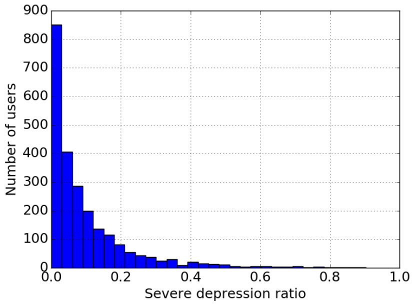


Figure 6: Distribution of the severe depression ratio of users. A total of 84.6% of users experienced at least one severe depression day in the dataset.

5.2 Labeled Dataset Creation

We defined *severe depression day* and assigned labels based on individual histories. A severe depression day is a day when the user experienced *negative feelings* (i.e., depressed or irritated) all day AND user was *physically inactive* (i.e., do nothing at home or sick in bed.) The distribution of the severe depression flags of the users is shown in Figure 6.

We used the dataset collected by the smartphone application described in Section 3. Successive histories of 28 days for each user were extracted as *instances*. User data do not overlap in the histories to avoid the unexpected *leakage* of the test dataset into a training dataset. We set the size as 28 to conduct different experiments (i.e., $n = 1, 3, 7$ and $k = 14, \ldots, 21$.) in a consistent manner.

For label extraction, we assigned severe depression labels for each block. Please note that the label extraction method differs for each severe depression task. For the $(1, 14)$-severe depression forecasting task, the 8th slot to the 21st slot were used to extract features and the 22nd slot was used for label extraction. For $(n, k) = (3, 14)$, we used the 22nd, 23rd, and 24th slots to calculate a severe depression label for the instances. This manner ensured that we used the same histories for different settings of n and k for comparative study. The dataset description is shown in Table 2.

5.3 RQ 1: Can we forecast severe depression?

We conducted a comparative experiment to evaluate the forecasting performance of LSTM-RNN to answer our RQ1.

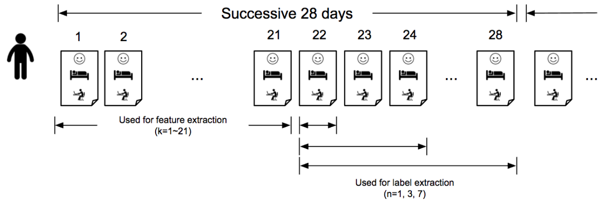


Figure 7: Instance extraction from a user's histories. Each block (instance) contains successive 28 days without overlapping with other instances.

Table 2: Dataset description.

Total days	345,158
Number of users	2,382
Avg. days / user	144.9
Number of severe / Nonsevere labels	32,205 / 312,953

We compared two different feature sets. The first feature set `severe only` employs the user's severe depression history to forecast severe depression. The feature set can be regarded as containing only minimal information of individual histories, assuming that a user only inputs his or her mood once a day by an aggregated manner. The other feature set `all` uses all the information of individual histories shown in Table 1 in addition to the day-of-the-week variable.

Our LSTM-RNN used three LSTM layers with 128 memory cells with dropout of 0.1, and a single fully-connected layer with 64 nodes. It was trained for 5 epochs using RMSProp [45].

We used SVM as a baseline learning algorithm because it has been used in previous studies for mood prediction [6][16][38]. Because SVM cannot directly handle time series data, we concatenated time series feature values to create a single feature vector, allowing SVM to learn a predictive model based on the same information. We used the (`all`) feature set to learn SVM. C parameter was selected from $\{10^{-3}, 10^{-2}, \ldots, 10^{1}\}$ by grid search.

For the experiment, we evaluated the methods for different n values ($n = 1, 3, 7$) and a fixed k value ($k = 14$) for the (n, k)-day severe depression forecast task. We used AUC-ROC as an evaluation metric because it is insensitive to imbalanced class distributions [11]. A random guessing classifier achieves 0.5 AUC-ROC. We conducted five-fold cross validation to calculate the evaluation measure. In this experiment, the cross validation split the dataset into training and test datasets on a user basis; The strategy ensures to evaluate the forecasting performance for unseen users whose histories are not used to train a predictive model in each fold.

The results are shown in Table 3. Remarkably, every classifier achieves significantly higher than random guessing (AUC-ROC=0.5). The results confirm the feasibility of forecasting severe depression based on individual histories. LSTM-RNN (`all`) outperforms SVM (`all`) regardless of n values. From the results, we confirm that LSTM-RNN appropriately processes time series information to forecast severe depression. Compared by feature sets, `all` shows higher AUC-ROC than `severe only`. Therefore, we confirm that fine-grained information such as moods at different parts of the day and action type inform the forecasting of severe depression.

As n increases from 1 to 7, AUC-ROC values of all the methods become lower. This indicates that it is difficult to forecast n-day severe depression with a large n value compared with a small n value. This follows our intuition that forecasting a longer future is more difficult than forecasting a shorter future. The general trend of the results is consistent over different n in Table 3.

We calculated the feature importance of LSTM-RNN (`all`) based on *Mean Decrease AUC-ROC*, which is the decrease value of AUC-ROC on the test dataset when the values of a selected feature are randomly permuted among all instances. This is commonly used for calculating feature importance, especially for ensemble classifiers [46] because it is not straightforward to evaluate feature importance for nonlinear classifiers including LSTM-RNN.

Figure 8 contains three figures representing feature importance distributions for different n values ($n = 1, 3, 7$). Figure 8(a) and Figure 8(b) present evening mood, which contributes to forecasting performance. For one-day forecasting, the most recent mood is the most informative signal to forecast severe depression for the coming day. On the other hand, the feature importance of seven-day forecasting shows that morning mood contributes most greatly to the predictive model. This indicates that morning moods in individual histories are more indicative of severe depression in the distant future. The results indicate that the day-of-the-week variable contributes to the forecasting performance well for one-day forecasting but not for the others. We believe the reason for this rests on the current definition of n-day severe depression, which does not point to a single day in the future if $n > 1$. Thus, the predictive model cannot utilize the day-of-the-week information.

Remarkably, sleeping features and medical features generally do not show a positive contribution to the forecasting performance. 1-day forecasting is the only case in which both sleeping hours and sleeping irregularity show a positive contribution. Medical features and nocturnal awakening do not show a positive contribution to the forecasting performance in any case. These results are inconsistent with previous studies that have shown a correlation between sleeping irregularity and negative moods [37]. We consider the absence of contribution by these features is caused by building a single predictive model among all users in the dataset. Some users might suffer from severe depression without attending hospitals or taking medicines. This result thus implies that stratifying users to create multiple predictive models for each group might improve forecasting performance.

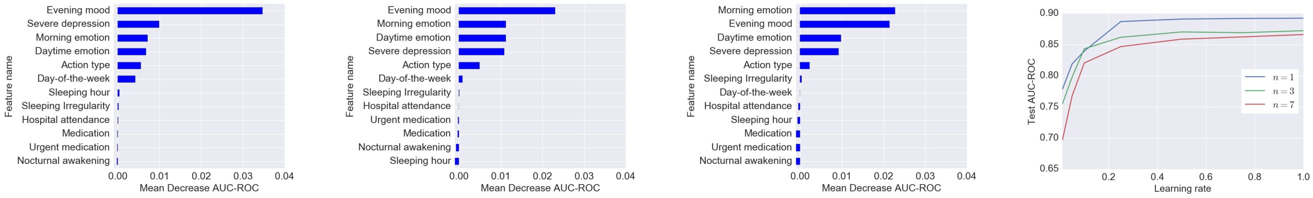
Figure 8(d) presents the learning curves of LSTM-RNN (`all`) for different n settings. Three curves consistently improve their performance as the ratio increases. The result verifies the appropriate learning of severe depression forecasting models by LSTM-RNN. The improvements saturate around 30% of training data in three models. This indicates that we need a large-scale dataset of about 700 users to utilize LSTM-RNN for depression forecast.

5.4 RQ 2: How many days should we look back?

To answer our RQ2, we evaluated our model with different k ($k = 1, 2, \ldots, 21$) to compare the forecasting performance

Table 3: Experimental results. The mean value of Test AUC-ROC is listed for each method. The standard deviations are in parentheses. *** denotes $p < 0.01$. The p-values were calculated based on paired t-test after making adjustments using the Holm-Bonferroni method for multiple comparison.

	AUC-ROC		
Method (`feature set`)	$n = 1$	$n = 3$	$n = 7$
SVM (`all`)	.837 (.031)	.822 (.045)	.822 (.047)
LSTM-RNN (`severe only`)	.846 (.032)	.821 (.042)	.800 (.053)
LSTM-RNN (`all`)	**.886***** (.020)	**.860*****(.031)	**.842*****(.044)



(a) (b) (c) (d)

Figure 8: (a) Mean Decrease AUC-ROC feature importance for each feature. (a) $n = 1$), (b) $n = 3$, and (c) $n = 7$. (d) Learning curves ($n = 1, 3, 7$).

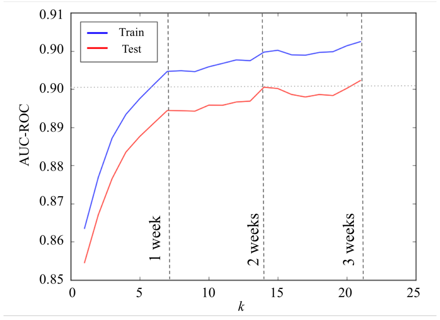


Figure 9: Forecasting performance over different days used for training. The x-axis denotes the number of days used for the model and the y-axis denotes the AUC-ROC values.

of predictive models. Here, $k = 1$ means that a predictive model uses only the day before a target day for feature extraction, and the $k = 21$ setting uses data from the previous three weeks for feature extraction. If information in the last several weeks contributes to forecasting performance, the performance should keep improving as k increases. We used LSTM-RNN (`all`) and fixed n as 1 for the experiment.

Figure 9 summarizes the results. The performance curve shows a rapid increase in the first week and a slower increase in the second and third weeks. The improvement apparently saturates as k increases. After k is higher than 14, the performance curve drops until k reaches 21. The results indicate that the history of the previous two weeks is sufficient to forecast future severe depression. This finding coincides with the design of commonly used mental health assessment such as the PHQ-9 and GAD-7, whose first questions are, "Over the last two weeks, how often have you been bothered by any of the following problems?" This consistency empirically ensures that the assessments' questions are reasonable for capturing individual mental health status.

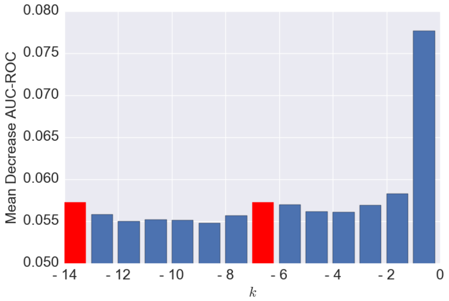


Figure 10: Feature importance distribution of different days in last k days used for severe depression forecasting. Mean Decrease AUC-ROC was used to calculate the feature importance. Higher values present a more important contribution to forecasting performance. $k = 7, 14$ are the same day of the week as the target day and are colored by red.

To precisely analyze the contribution of every single day in the histories of the previous two weeks, we also calculated the importance of each day of the last k days based on Mean Decrease AUC-ROC in the same manner as the feature importance analysis in 5.3. Instead of selecting a single feature, this analysis selected a single day from last k days to randomly permute the feature values correspond to the selected day. This analysis shows the contribution of a day in last k days for depression forecasting performance.

The results are shown in Figure 10. This figure indicates that user logs from the most recent day are most informative features for depression forecasting. This follows the intuition that a present mental status strongly depends on the previous state. However, user logs from other days also have reasonably high values. The finding is consistent with the results in Figure 9. The implications here are twofold: (1) there exist long-range dependencies in depression forecasting with user's historical logs, and (2) the long-range dependen-

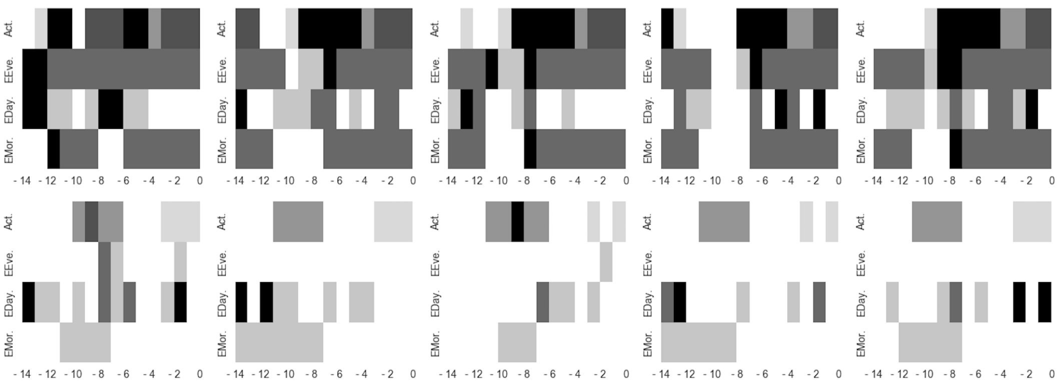


Figure 11: Ten representative high-level representations extracted from the hidden layer of the trained LSTM-RNN. Rows of each picture denote (1) mood in the morning (EMor), (2) in the afternoon (EDay), (3) in the evening (EEve), and (4) action type (Act). Columns denote the days of input information. The darker color represents more depressed mood. (a) Five representations in the above row are selected for the severe depression class, and (b) the other five in the below row are selected for the non-severe depression class.

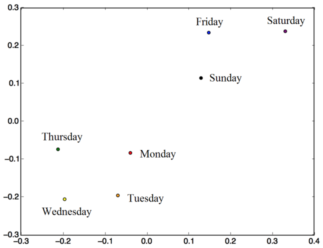


Figure 12: Categorical variables of days of the week in embedding space.

cies can be expressed by a finite number of patterns among all the users so that LSTM-RNN performs well in forecasting severe depression in the test dataset. Assumption (1) is also supported by the results in Figure 9. Interestingly, a present mental status depends on not only the most recent day but also other days. Long-range dependency is especially apparent if the day is the same day of the week as the target day.

To confirm the learned representation of LSTM-RNN, we looked up the representative input representation for each node in the fully-connected layer. We restored input representations that maximally activate a node in order to interpret the obtained high-level representations. The extracted patterns are shown in Figure 11. The representative patterns of the severe depression class include negative feelings (depressed or irritated) and inactive behaviors (do nothing at home, sick in bed), whereas the nonsevere depression class is associated with more positive feelings and more active behaviors in general. However, by focusing on moods in the afternoon (EDay in Figure 11), we confirm that representative patterns of the severe depression class still include positive feelings (i.e., fine or fair.) Additionally, the representative patterns of the nonsevere depression class include some negative feelings. These remarkable results do not match our intuition and indicate that negative feelings in the afternoon do not always indicate future severe depression.

As another example of interpretability of our method, we show the trained embedding vector of the day-of-the-week variable. Figure 12 presents the embedding vectors in two-dimensional space after converting original embedding vectors into two-dimensional vectors by singular value decomposition (SVD) [19]. Two groups are shown in the figure. Friday, Saturday, and Sunday are happier days than the other days, namely weekdays except Friday. This result follows the findings by this paper and previous studies. The model only uses the severe depression labels as target values to infer the day-of-the-week embedding representation.

6. CONCLUSION

This study tackled a novel type of depression prediction task—namely, depression forecast based on individual histories. Our study used 345,158 total days of logs collected from more than 2,382 users over 22 months. With the large-scale data, we utilized the power of deep learning to build a depression forecasting model. Our model separately embedded categorical variables, including our proposed day-of-the-week variable.

Experimental results confirmed that our framework was able to forecast severe depression based on individual histories with high accuracy. The results showed that fine-grained information such as reporting moods in different parts of the day improved forecasting performance. The capability of LSTM-RNN to incorporate long-range dependencies of time series helped us determine the contribution of distant past histories up to the previous two weeks. The representative patterns of the model further showed that having a depressed mood only in the afternoon is not always a sign of future severe depression.

This study relied on self-reported histories. However, previous studies have used many methods to estimate individual behavioral patterns, including sleeping and moods, based on behavioral and mobility information collected by smartphones and/or wearable devices. As we have shown in Figure 1, our aim is to bridge the gap between previous studies and depression forecasting. We believe that this paper is a good step toward automatic depression detection, which could provide appropriate interventions for people and improve well-being in our society.

7. REFERENCES

[1] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland. Social fMRI: Investigating and shaping social mechanisms in the real world. volume 7, pages 643–659, 2011.

[2] P. R. Aylard, J. H. Gooding, and P. J. McKenna. A validation study of three anxiety and depression self-assessment scales. *Journal of Psychosomatic Research*, 31(2):261–268, 1987.

[3] D. Ben-Zeev, E. A. Scherer, R. Wang, H. Xie, and A. T. Campbell. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric Rehabilitation Journal*, 38(3):218–226, 2015.

[4] A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. Pentland. Pervasive stress recognition for sustainable living. pages 345–350, 2014.

[5] M. N. Burns, M. Begale, J. Duffecy, D. Gergle, C. J. Karr, E. Giangrande, and D. C. Mohr. Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research*, 13(3):e55–17, 2011.

[6] L. Canzian and M. Musolesi. Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility trace analysis. In *Proc. UbiComp '15*, pages 1293–1304, 2015.

[7] P. Chow, H. Xiong, K. Fua, W. Bonelli, and B. A. Teachman. SAD: Social anxiety and depression monitoring system for college students. pages 125–130, 2016.

[8] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[9] A. A. Farhan, J. Lu, J. Bi, A. Russell, B. Wang, and A. Bamis. Multi-view Bi-clustering to Identify Smartphone Sensing Features Indicative of Depression. In *Proc. IEEE CHASE '16*, pages 264–273, 2016.

[10] A. A. Farhan, C. Yue, R. Morillo, S. Ware, and J. Lu. Behavior vs. introspection: Refining prediction of clinical depression via smartphone sensing data. *Proc. IEEE Wireless Health Conference*, 2016.

[11] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[12] P. L. Gerald B. Can Smartphones Detect Stress-related Changes in the Behaviour of Individuals? In *Proc. PerCom '12*, pages 423–426, 2012.

[13] S. A. Golder and M. W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[15] Y. Huang, H. Xiong, K. Leach, Y. Zhang, P. Chow, K. Fua, B. A. Teachman, and L. E. Barnes. Assessing social anxiety using gps trajectories and point-of-interest data. In *Proc. UbiComp '16*, pages 898–903, 2016.

[16] N. Jaques, S. Taylor, A. Azaria, A. Ghandeharioun, A. Sano, and R. W. Picard. Predicting students' happiness from physiology, phone, mobility, and behavioral data. In *Proc. ACII*, 2015.

[17] N. Jaques, S. Taylor, and A. Sano. Multi-task, Multi-Kernel Learning for Estimating Individual Wellbeing. *Proc. NIPS MultiML Workshop*, 2015.

[18] T. R. Kirchner and S. Shiffman. Spatio-temporal determinants of mental health and well-being: advances in geographically-explicit ecological momentary assessment (GEMA). *Social Psychiatry and Psychiatric Epidemiology*, 51(9):1211–1223, 2016.

[19] V. Klema and A. Laub. The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control*, 25(2):164–176, 1980.

[20] K. Kroenke, R. L. Spitzer, and J. B. W. Williams. The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613, 2001.

[21] N. D. Lane, M. Lin, M. Mohammod, X. Yang, H. Lu, G. Cardone, S. Ali, and A. Doryab. BeWell: Sensing Sleep, Physical Activities and Social Interactions to Promote Wellbeing. *Mobile Networks and Applications*, 19(3):1–15, 2014.

[22] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong. MoodScope - building a mood sensor from smartphone usage patterns. In *Proc. MobiSys '13*, 2013.

[23] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell. Learning to Diagnose with LSTM Recurrent Neural Networks. *arXiv.org*, Nov. 2015.

[24] A. L. Lopresti, G. L. Maker, S. D. Hood, and P. D. Drummond. A review of peripheral biomarkers in major depression: The potential of inflammatory and oxidative stress biomarkers. *Prog Neuropsychopharmacol Biol Psychiatry*, 48:102–111, 2014.

[25] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury. StressSense - detecting stress in unconstrained acoustic environments using smartphones. pages 351–360, 2012.

[26] Y. Ma, B. Xu, Y. Bai, G. Sun, and R. Zhu. Daily mood assessment based on mobile phone sensing. In *Proc. BSN '12*, pages 142–147, 2012.

[27] G. MacKerron and S. Mourato. Happiness is greater in natural environments. *Global Environmental Change*, 23(5):992–1000, 2013.

[28] A. Madan, M. Cebrian, D. Lazer, and A. Pentland. Social sensing for epidemiological behavior change. In *Proc. Ubicomp '10*, pages 291–300, 2010.

[29] M. Matthews, S. Abdullah, G. Gay, and T. Choudhury. Tracking Mental Well-Being - Balancing Rich Sensing and Patient Needs. *IEEE Computer*, (4):36–43, 2014.

[30] A. Mehrotra, R. Hendley, and M. Musolesi. Towards multi-modal anticipatory monitoring of depressive states through the analysis of human-smartphone interaction. In *Proc. UbiComp '16*, pages 1132–1138, 2016.

[31] D. S. Moskowitz and S. N. Young. Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology.

Journal of Psychiatry & Neuroscience, 31(1):13–20, 2006.

[32] S. T. Moturu, I. Khayal, N. Aharony, W. Pan, and A. Pentland. Using social sensing to understand the links between sleep, mood, and sociability. In *Proc. SocialCom/PASSAT*, pages 208–214, 2011.

[33] M. Rabbi, S. Ali, T. Choudhury, and E. Berke. Passive and In-Situ assessment of mental and physical well-being using mobile sensors. *UbiComp*, pages 385–394, 2011.

[34] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas. Emotionsense - a mobile phones based adaptive platform for experimental social psychology research. In *Proc. UbiComp '10*, pages 281–290, 2010.

[35] A. Raiker, J. Latayan, S. Pagsuyoin, and A. Mathieu. Use of biomarkers in depression diagnostics. In *Proc. SIEDS*, pages 245–249, 2016.

[36] 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. *Journal of Medical Internet Research*, 17(7):e175–11, 2015.

[37] A. Sano. *Measuring College Students' Sleep, Stress, Mental Health and Wellbeing with Wearable Sensors and Mobile Phones*. PhD thesis, MIT, 2015.

[38] A. Sano, A. J. K. Phillips, A. Z. Yu, A. W. McHill, S. Taylor, N. Jaques, C. A. Czeisler, E. B. Klerman, and R. W. Picard. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. In *Proc. BSN '15*, pages 1–6, 2015.

[39] A. Sano and R. W. Picard. Stress Recognition Using Wearable Sensors and Mobile Phones. In *Proc. ACII '13*, pages 671–676, 2013.

[40] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe. A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine*, 166(10):1092–1097, 2006.

[41] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and the Patient Health Questionnaire Primary Care Study Group. Validation and Utility of a Self-report Version of PRIME-MD: The PHQ Primary Care Study. *JAMA*, 282(18):1737–1744, 1999.

[42] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:929–1958, 2014.

[43] A. A. Stone and S. Shiffman. Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine*, 1994.

[44] M. P. Taylor. Tell me why I don't like Mondays: investigating day of the week effects on job satisfaction and psychological well-being. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(1):127–142, 2006.

[45] T. Tieleman and G. Hinton. Lecture 6.5 - RMSProp, COURSERA: Neural networks for machine learning. Technical report, 2012.

[46] E. Tuv, A. Borisov, G. Runger, and K. Torkkola. Feature selection with ensembles, artificial variables, and redundancy elimination. *Journal of Machine Learning Research*, 10:1341–1366, 2009.

[47] P. Wang, J. Guo, Y. Lan, J. Xu, and X. Cheng. Your cart tells you: Inferring demographic attributes from purchase data. In *Proc. WSDM '16*, pages 173–182, 2016.

[48] R. Wang, M. S. H. Aung, and S. Abdullah. 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*, pages 886–897, 2016.

[49] 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 *Proc. UbiComp '14*, pages 3–14, 2014.

[50] D. Zhou, J. Luo, V. M. B. Silenzio, Y. Zhou, J. Hu, G. Currier, and H. A. Kautz. Tackling Mental Health by Integrating Unobtrusive Multimodal Sensing. In *Proc. AAAI '15*, pages 1401–1408, 2015.