Improving Students' Daily Life Stress Forecasting using LSTM Neural Networks

Terumi Umematsu*†, Akane Sano‡*, Sara Taylor*, Rosalind W. Picard*
*Massachusetts Institute of Technology, Media Lab
75 Amherst Street, Cambridge, MA, USA, 02139
{terumi,sataylor,picard}@media.mit.edu, akane.sano@rice.edu
†NEC Corporation, Biometrics Research Laboratories
‡Rice University, Department of Electrical and Computer Engineering

***Abstract*—Accurately forecasting stress may enable people to make behavioral changes that could improve their future health. For example, accurate stress forecasting might inspire people to make changes to their schedule to get more sleep or exercise, in order to reduce excessive stress tomorrow night. In this paper, we examine how accurately the previous N-days of multi-modal data can forecast tomorrow evening's high/low binary stress levels using long short-term memory neural network models (LSTM), logistic regression (LR), and support vector machines (SVM). Using a total of 2,276 days, with 1,231 overlapping 8-day sequences of data from 142 participants (including physiological signals, mobile phone usage, location, and behavioral surveys), we find the LSTM significantly outperforms LR and SVM with the best results reaching 83.6% using 7 days of prior data. Using time-series models improves the forecasting of stress even when considering only subsets of the multi-modal data set, e.g., using only physiology data. In particular, the LSTM model reaches 81.4% accuracy using only objective and passive data, i.e., not including subjective reports from a daily survey.**

***Index Terms*—Stress, Forecasting, Objective, Wearable, LSTM**

I. Introduction

Stress is a complex and dynamic process that can help a person in many beneficial ways to successfully confront a challenge or threat, but also can cause a negative emotional response when an individual feels that the environmental demands exceed their adaptive capacity [1]. Researchers have shown that stress increases susceptibility to infection and illness [2] and affects a diverse range of physical, psychological and behavioral conditions, such as anxiety, depression, and sleep disorders [3]. In addition, excessive stress, which is widespread in today's society, can decrease job productivity and negatively affect overall well-being [4].

The ability to forecast stress levels could enable better self-management of one's behavioral choices in ways that might prevent excessive stress. The ability to model and forecast stress could be immensely beneficial, especially if such a forecast could be made using data collected in an unobtrusive and privacy-sensitive way. In this work, we show that wearable sensors and mobile phones, coupled with machine learning, can provide a significantly accurate forecast of future stress.

This work is supported by National Institute of Health (R01GM105018), the MIT Media Lab Consortium, NEC Corporation and Samsung Electronics.

Some technologies have been developed not only to estimate current human well-being levels but also to forecast them [5], [6]. Suhara et al. have shown that a depressed mood can be predicted with a high accuracy using self-reported survey data from the past several days [5]. This work used LSTM to predict mood given two weeks of mood history reported daily, learning the function $p(y_{t+1}|y_1, ..., y_t, z_1, ..., z_t)$, where z_t is all the data collected from behavioral surveys about a person on day t, and y_{t+1} is a self-reported depressed mood the following day. Using a large-scale dataset of 2,382 people and a total of 345,158 days, they achieved an Area Under the Curve of the Receiver Operating Characteristic (AUC-ROC) of 0.886 forecasting severely depressed mood or not.

Taylor et al. used personalized multi-tasking neural network models to predict tomorrow's well-being (good/poor mood, good/poor health and high/low stress) given multi-modal data from today ($p(y_{t+1}|x_t)$), obtaining accurate predictions about an individual's next-day mood through personalization, without requiring previous self-reported labels (e.g., y_t) [6]. The results have shown that tomorrow's mood, stress and health levels can be predicted with 78-82% accuracy using a personalized model based on today's physiological, mobile phones, and behavioral survey data. However, it has not yet been examined whether stress forecasting accuracy can be improved using the previous several days of only passively collected data (i.e. physiological and mobile phone data), or whether significant accuracy can be achieved without personalization. Taylor et al.'s best results without personalization achieved 68% accuracy on forecasting tomorrow's stress.

In this paper, we evaluated the stress forecasting accuracies from static models (support vector machine (SVM) and logistic regression(LR)) and time-series models (long short-term memory neural network models (LSTM)) using the previous 1 to 7 days of physiological, mobile phone, and behavioral survey data. Given that current and future human mental conditions are affected by the past few days [2], we hypothesize that an LSTM model, which can exploit long-term memory, will improve the forecasts.

This paper contains two main contributions: First, we forecast future stress using data collected from college student participants (N = 142) in their daily lives, using wearable sensors, mobile phones, and behavioral surveys; hence, the

LSTM model we build is directly relevant to daily-life stress forecasting. Second, we found that, without requiring any personalization, the LSTM model significantly outperforms the LR and SVM with the best results reaching 83.6% accuracy using 7 days of prior data to forecast tomorrow's binary high/low stress level. Furthermore, it can forecast tomorrow's stress with high accuracy (81.4%) using only objective and passive data (i.e. data sensed directly from wearable sensors and mobile phones). Thus, people do not have to be interrupted to fill out surveys each day. These two contributions make this work a valuable step towards developing a practical stress forecasting system.

II. Data

A. Dataset and Classification Labels

The data in this experiment were collected in a study to measure Sleep, Networks, Affect, Performance, Stress, and Health using Objective Techniques (SNAPSHOT) [7]. The study gathered 30-day multi-modal data, including physiological, mobile phone, and behavioral survey data from college students at a US university. The study participants obtained compensation based on their contribution to the study. Stress scores were collected every morning and evening using self-reported scores from 0 (stressed out) - 100 (calm).

We framed the problem as a binary classification; days on which a participant reported a stress-calm score in the top 40% of all stress-calm scores are labeled as a low-stress day, and days in which participants reported a stress-calm score in the bottom 40% are labeled as a high-stress day. We discarded only the middle 20% of scores similar to Taylor et al. [6]. In this paper, we used a total of 1,231 sequences of 8 consecutive days of data from 142 participants (these 8-day-long sequences are overlapping resulting in a total use of 2,276 days of data).

B. Feature Calculation

We computed 375 daily features including 37 behavioral survey (excluding self-reported stress scores), 173 physiology, 150 mobile phone, and 15 mobility features. The feature modalities are explained in detail below.

(a) Survey:

Participants filled out a survey about their daily behaviors every evening. They self-reported the timing and duration of a variety of activities, including sleep, academic activities, extracurricular activities, and exercise. Whether the participant engaged in social activity before bed, amount of caffeine intake, and whether a positive or negative social interaction was experienced were also self-reported.

(b) Physiology:

The physiological measurements collected by wrist-worn Affectiva Q sensors at 8 Hz include 24-hour-a-day electrodermal activity (EDA) measured as skin conductance (SC), skin temperature (ST), and 3-axis accelerometer. Features such as step count, stillness, and SC responses were calculated — all of which relate to emotional arousal and stress. EDA, acceleration and ST were collected to measure sympathetic nervous activity, physical activity, sleep patterns, circadian rhythm, and stress responses [8]–[10]. Following [11] and [6], for each time span (all-day, midnight-3am, 3am-10am, 10am-5pm, 5pm-midnight) the following sets of features were computed: EDA Peak features (for all detected peak features and for only non-artifact peaks [12]), SC level features, accelerometer features, temperature features, and various combinations of the three physiological data streams.

(c) Phone:

The phone log data consist of information about the timing, type, and duration of phone calls and SMS messages, and times the screen was turned on and off. We assumed that there are two main mechanisms through which screen logs and communication information can affect wellbeing: (1) light from the screen can disrupt circadian rhythms and therefore sleep [13], and (2) the amount of social support in a person's life is strongly linked to resilience to depression [14], [15]. As with physiology, the features were computed over the time intervals defined in the previous setion (II.B(b)) spanning the course of the day.

(d) Mobility:

In addition to communication and screen events, the phone app logged the participants' GPS coordinates throughout the day, as well as whether they were using Wifi or cellular data. Previous studies have shown that mobility patterns are linked with mental health states [16], [17]. We followed the method of Jaques et al. [18] to down-sample the signal and compute features such as the total distance traveled, statistical features about distance traveled in 5 minutes, and the amount of time spent on campus.

III. Methods

We conduct a series of experiments to examine whether we can improve the stress forecasting accuracy using non-personalized temporal machine learning models.

A. Long Short Term Memory Networks (LSTM)

LSTMs [19] have the ability to learn long-term dynamics while avoiding vanishing and exploding gradient problems and have recently gained great success in sequence learning tasks such as speech recognition and machine translation. We designed our LSTM with a single LSTM layer with 32 nodes and a dropout of 0.2. Drop-outs were used between the LSTM and dense layers. The output of the last cell of the LSTM layer was connected to a dense layer. Finally, a sigmoid activation layer predicted the high/low stress levels. We trained our LSTM using RMSprop [20] with binary cross-entropy loss and an iteration number of 1000. The whole algorithm was implemented using deep learning frameworks Keras 2.1.3 and Python 3.5.4. Fig. 1 presents an overview of our method.

B. Static Methods

For comparison to the LSTM, we used an SVM classifier with a radial basis function kernel and LR, because these static methods are widely recognized as performing well, and were used in previous studies for mood prediction [5], [6], [18]. Because SVM and LR cannot directly exploit temporal data,

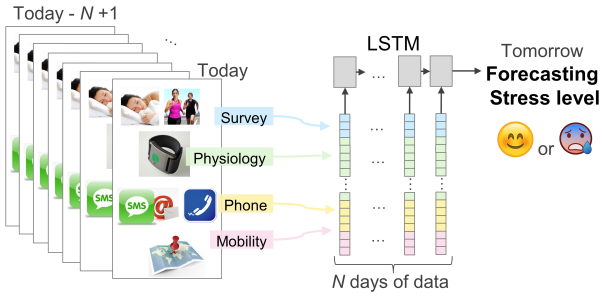

Fig. 1. Overview of our method

we concatenated the time series feature values to create a single feature vector, allowing SVM and LR to learn a forecast model based on the same time-series information.

C. Experiments

We examine how accurately the previous N-day multimodal data can forecast a "tomorrow" high/low (binary) stress level using time-series (LSTM) and two static models (SVM and LR). We used accuracy and AUC-ROC as evaluation metrics. A baseline (random) classifier achieved 54.1% accuracy (high/low stress data: 565/666).

The full dataset (1,231 sequences) was then used in a five-fold cross validation with 80% of the data for training and validating the models, and 20% for testing for each fold. While it's possible days could have been repeated within the train/validation loops, the days in the test set were kept completely independent of the train and validation data. Specifically, within the training and validation set, we used 80% of the dataset for training and 20% as validation and selected the hyperparameters (LSTM: dropout and iteration number, SVM: C and gamma, LR: C) that yielded the highest accuracy on the validation set. We calculated the average and the standard deviation of the test set for the 5 folds when reporting evaluation metrics.

Using the set up above, we conducted the following two experiments:

(1) Forecasting tomorrow's stress using all features: We evaluated the three methods using different lengths of previous history to forecast next-day stress. Specifically, the model learned $p(y_{t+1}|x_t, ..., x_{t-N+1})$, the probability of the person's stress tomorrow given various N of the previous days (N = 1 - 7), where x_t is all the data collected from behavioral surveys, wearable sensors, and mobile phones on day t, and y_{t+1} is the binary self-reported stress label the evening of the following day. We also retrained models and computed metrics for each model using only a single modality (e.g., survey only).

(2) Forecasting tomorrow's stress using only objective data: In this experiment, we evaluated the three methods using the previous N days of only the objective data in forecasting tomorrow' stress. In other words, we excluded the 37 survey features and used only the 338 objective features from wearables and mobile phones to forecast tomorrow's binary stress level.

IV. Results

(1) Forecasting tomorrow's stress

The stress forecasting accuracies of the SVM, LR and LSTM models using 1-7 days of data are shown in Fig. 2. The results show that accuracy is improved by using the LSTM instead of the static models (with LSTM accuracy $>$ SVM $>$ LR) and by using a longer history in each modality (p $<$0.05, Analysis of variance (ANOVA) and Tukey's honest significant difference (HSD) test). In LSTM, the model with 7 days of data showed significantly higher accuracy than the model with only 1 day of data. In particular, we found that 4 days of data in the LSTM model is the smallest number of days required to be significantly better than using only 1 day of data. This finding also holds when considering the four different modalities (survey, physiology, phone, and mobility) individually.

The best results overall were obtained using the LSTM with all features (including the survey features), 83.6% accuracy (AUC-ROC 0.831) using the previous 7 days. In addition, we also evaluated a participant-independent model using all features, and the accuracies are statistically similar to the participant-dependent models (p>0.05, ANOVA, for N = 1, 2, 3, 5, and 7 days). Furthermore, we also evaluated forecasting binary stress levels one week in the future, and obtained 80.1% using data of the previous 7 days from today.

Nearly the same performance was achieved using the LSTM with data only from the single modality of physiology, with 82.4% (AUC-ROC 0.820) using the previous 7 days of physiology features. This performance is not significantly different ($p = 0.609$, ANOVA) than using all features, while it requires only a wearable sensor to collect the data. The improvement of the LSTM time-series models over the static models was most dramatic for the physiology features where the LSTM was significantly better than the LR, which was significantly better than the SVM. (For other cases, the SVM was usually better than LR.)

We computed the top 10 of the mean absolute weights of each feature across all connected nodes in the input layer of the LSTM using all 7 days of features and all feature modalities. Features with higher weights indicate a stronger influence on forecasting stress. The top ten features were (1) whether or not the participant slept (from the survey), (2) if the participant had a memorable negative social interaction (survey), (3) a mobility feature (distance within 5 minutes), (4) EDA peaks during 3-10H, (5) EDA peaks during 17-24H, (6) missed call timestamps of 0-24H, (7) EDA peaks during 10-17H, (8) the number of minutes the participant spent awake after falling asleep at night (survey), (9) overslept or not (survey), and (10) EDA peaks during 0-3H. Therefore, all four of the modalities were represented in the top 10 weighted features.

As we can see in Table I, the LSTM is relatively balanced in misclassifying tomorrow as a stressed day when it is actually calm and vice-versa. The static models, on the other hand, are more imbalanced and tend to classify tomorrow as a calm day even when it should be labeled as stressed. Specifically,

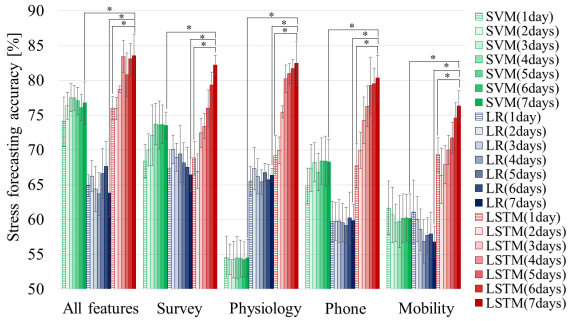

Fig. 2. Stress forecast accuracy for support vector machine (SVM), logistic regression (LR), and long short-term memory neural network models (LSTM) (* : One way ANOVA, Tukey's HSD test, p <0.05)

TABLE I
CONFUSION MATRIX FOR THE 7-DAY ALL FEATURE MODELS

		Predicted Label					
		LSTM		**SVM**		**LR**	
		Stress	Calm	Stress	Calm	Stress	Calm
Actual Label	Stress	438	127	334	231	152	413
	Calm	75	591	55	611	32	634

54.1% of the days in our data set are calm; the LSTM model labels a total of 58.3% of days as calm, whereas the SVM and LR models label 68.4% and 85.1% of the days as calm. These errors by the static models are particularly undesirable if a goal of a supportive system is to help people take action in preventing a forecasted stressed day from being so stressful; the SVM and LR would miss many opportunities to help.

(2) Stress forecasting using only objective data

When using only the objective features the best accuracy for tomorrow night's stress, 81.4% (AUC-ROC 0.809), was obtained with 7 days of data, which was found to be not significantly different ($p = 0.233$, ANOVA) than using all of the features. Thus, using LSTM maintained the same accuracies without survey data as with adding survey data. However, if survey data is not used in the SVM model, the stress forecasting accuracy decreases significantly by about 5-9% ($p < 0.05$, ANOVA and Tukey's HSD test). The best results obtained using the LR with only objective features were 63.5% accuracy (AUC-ROC 0.608) using the previous 1 day, and for the SVM were 69.4% accuracy (AUC-ROC 0.671) using the previous 1 day. Therefore, we can forecast subjective stress level more accurately using only objective features when we use the LSTM. High accuracy forecasting of stress using data collected in an unobtrusive way, such as using wearable sensors and mobile phones, is an effective way to monitor the stress in our daily life using only passive information; no extra time or effort is required from the participants beyond keeping their devices charged.

V. CONCLUSIONS

In this work, we applied LSTM to objective data including physiological signals, mobile phone usage, and mobility patterns, and to self-report data from behavioral surveys. We examined how accurately the previous N-day multi-modal data can forecast tomorrow's evening stress level using LSTM, LR, and SVM. Using a total of 2,276 days, with 1,231 overlapping 8-day sequences of data from 142 students, we found the LSTM outperforms the LR and SVM with the best results reaching 83.6% using 7 days of prior data. We confirmed that using LSTM improves the forecasting of stress across all the modalities, individually or ensemble. In addition, we made forecasts of tomorrow's stress with 81.4% accuracy based on using only objective and passive data such as physiology and phone data, without having to use survey data.

In the future, we plan to examine other LSTM, deep learning models and regression models to improve future stress forecasting accuracies. We will also examine how this model, and others, can help illuminate modifiable behavioral features (e.g., bedtime) that contribute, in an evidence-based way, to each individual's wellbeing so that the model can be used as a tool to help individuals improve their personal well-being.

REFERENCES

[1] S. Cohen, D. Janicki-deverts, and G. E. Miller, "Psychological Stress and Disease," vol. 298, no. 14, pp. 1685–1687, 2015.

[2] S. Cohen, D. A. Tyrrell, and A. P. Smith, "Psychological Stress and Susceptibility to the Common Cold," *New England Journal of Medicine*, vol. 325, no. 9, pp. 606–612, 1991.

[3] S. Cohen, R. C. Kessler, and L. U. Gordon, *Measuring stress: A guide for health and social scientists*. Oxford University Press on Demand, 1997.

[4] T. W. Colligan and E. M. Higgins, "Workplace stress: Etiology and consequences," *Journal of Workplace Behavioral Health*, vol. 21, no. 2, pp. 89–97, 2006.

[5] Y. Suhara, Y. Xu, and A. S. Pentland, "DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks." *World Wide Web Conference*, pp. 715–724, 2017.

[6] S. A. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health," *IEEE Transactions on Affective Computing*, no. 99, pp. 1–14, 2017.

[7] A. Sano *et al.*, "Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: Observational study," *J Med Internet Res*, vol. 20, no. 6, p. e210, Jun 2018.

[8] W. Boucsein, *Electrodermal activity*. Springer Science & Business Media, 2012.

[9] C. G. Scully *et al.*, "Skin surface temperature rhythms as potential circadian biomarkers for personalized chronotherapeutics in cancer patients," *Interface Focus*, vol. 1, no. 1, pp. 48–60, Feb. 2011.

[10] K. A. Herborn *et al.*, "Skin temperature reveals the intensity of acute stress," *Physiology & Behavior*, vol. 152, pp. 225–230, Dec. 2015.

[11] S. A. Taylor, "Characterizing Electrodermal Responses during Sleep in a 30-day Ambulatory Study," *MIT, Master's Thesis*, 2016.

[12] S. Taylor *et al.*, "Automatic identification of artifacts in electrodermal activity data," *EMBC*, pp. 1934–1937, 2015.

[13] C. A. Czeisler *et al.*, "Bright light resets the human circadian pacemaker independent of the timing of the sleep-wake cycle," *Science*, vol. 233, no. 4764, pp. 667–671, 1986.

[14] M. E. Seligman, "Flourish: a visionary new understanding of happiness and well-being," *Policy*, vol. 27, no. 3, pp. 60–61, 2011.

[15] R. S. Peirce, M. R. Frone, M. Russell, M. L. Cooper, and P. Mudar, "A longitudinal model of social contact, social support, depression, and alcohol use." *Health Psychology*, vol. 19, no. 1, pp. 28–38, 2000.

[16] L. Canzian and M. Musolesi, "Trajectories of depression," *UbiComp*, pp. 1293–1304, 2015.

[17] S. Saeb *et al.*, "Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study," *Journal of Medical Internet Research*, vol. 17, no. 7, p. e175, jul 2015.

[18] N. Jaques *et al.*, "Predicting students' happiness from physiology, phone, mobility, and behavioral data," *ACII*, pp. 222–228, 2015.

[19] S. Hochreiter and J. Urgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

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