Personalized Wellbeing Prediction using Behavioral, Physiological and Weather Data

Han Yu[1], Elizabeth B. Klerman[2], Rosalind W. Picard[3], Akane Sano[1]

***Abstract*—We built and compared several machine learning models to predict future self-reported wellbeing labels (of mood, health, and stress) for next day and for up to 7 days in the future, using multi-modal data. The data are from surveys, wearables, mobile phones and weather information collected in a study from college students, each providing daily data for 30 or 90 days. We compared the performance of multiple models, including personalized multi-task models and deep learning models. The best personalized multi-task linear model showed mean absolute errors of 12.8, 11.9, and 13.7 on a continuous-100 pt scale for estimating next days mood, health, and stress value, while the best multi-task neural network model, applied to 3-way high/med/low classification of the wellbeing values showed F1 scores of 0.71, 0.74, and 0.66 on mood, health, and stress metrics, respectively. We found that features related to weather, and morning academic activities are strongly associated with wellbeing labels. We further found greater prediction accuracy among participants with the least fluctuations in their wellbeing labels.**

***Index Terms*—Wellbeing Prediction, Personalized models, Multi-task Learning, LSTM, CNN, Regression, 3-class classification, Mood, Health, Stress, Wearables, Mobile phone**

I. Introduction

Wellbeing - the presence of positive emotions and moods and physical health [1]- is composed of multiple mental and physical factors that are usually measured with self-report surveys [2]. Since wellbeing is associated with health, productivity, and disease risks [3], studying wellbeing is important for individuals and society.

Research for measuring or predicting wellbeing has been conducted using objective physiological and behavioral sensors [4] [5]. We designed the SNAPSHOT study [6] to quantify physiological and behavioral factors related to human wellbeing using multi-modal data from surveys, wearable sensors and mobile phones. Our ultimate goal is predicting an individual's wellbeing trend and providing personalized warnings and interventions before wellbeing-related problems become severe. We have defined daily wellbeing using self-reported mood, health and stress on non-numeric (scored as 0-100) scales. Our previous work predicted these perceived wellbeing labels using machine learning [7] [8] and showed that multi-task learning based personalized regression [8] and binary (high/low; top and bottom 40 % of the scale) prediction models can predict the next day's self-reported mood, health, and stress using current and previous days' data [7]. Previous work also showed that using 7 days of time-series data with recurrent neural network (RNN) models can give acceptable results in wellbeing prediction without building personalized models [9]. These published studies have limitations, including: (i) They framed the prediction tasks as binary (keeping only the highest 40% and lowest 40% of the wellbeing labels and discarded all data in the middle 20% of the range.) (ii) They used RNN models that were designed for the overall dataset instead of learning representations for each individual, and (iii) They explored only a limited number of learning models: convolutional neural networks (CNN) have not yet been evaluated.

In this paper, we extend the prior work and develop personalized regression models for predicting self-reported mood, health and stress on a continuous non-numeric scale and personalized three-class classification models for predicting high/mid/low states of labels. We compare several machine learning algorithms, including multitask linear models, RNN models and CNN models to evaluate how well they predict future wellbeing labels based on time-series data from the past.

II. Methods

A. Dataset

We used the dataset collected in the SNAPSHOT study [6] that includes multi-modal physiological and behavioral data from 251 college students in one university (total 8430 days). The data recorded include gender, Big Five Personality scores [10], wrist wearable sensor data (acceleration, skin conductance and temperature), mobile phone data (call, sms, screen on/off logs, location), weather data (obtained using DarkSky API [11]), daily survey data (sleep, academic, exercise activities and social interactions), and self-reported daily evening wellbeing (non-numeric scales of mood, health and stress later scored 0-100). Between 2013-2017, consecutive 30-day data were recorded for 236 participants and consecutive 90-day data were recorded for 15 participants.

We computed 420 features including timing and duration of calls, sms, and screen usage; mobility patterns (radius, distance); number of steps; skin temperature and conductance responses; self-reported caffeine and drug intake; academic and exercise activity timing and duration; and weather metrics.

B. Multi-Task Learning

Multi-tasking learning (MTL) can optimize multiple related tasks together [12] by sharing the information and representations in the learning process. The generalization effect of the

This work was supported by NSF(#1840167), NIH (R01GM105018, K24-HL105664), Samsung Electronics, and NEC Corporation. We thank our SNAPSHOT study collaborators and study participants.

[1]Department of Electrical and Computer Engineering, Rice University. hy29@rice.edu

[2]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Harvard Medical School.

[3]Media Lab, Massachusetts Institute of Technology.

associated MTL is usually superior to that of the single task learning [13].

Our data include similarities and differences among participants. One condition for study participation was knowing other people in the study. Therefore, some participants in the study may have similar study, sleep and exercise schedules as well as the usual commonalities across college students. Therefore, in this paper, we applied MTL to data from (i) different participants (persons as tasks) and (ii) different groups of participants clustered using Gaussian mixture models for Big Five personality types and gender (a group of people as tasks) as related tasks. The number of clusters (groups) was finalized by the highest silhouette scores.

We compared the multiple algorithms that included interpretable linear models as well as neural network models that are hard to interpret. We describe the algorithms that we used in the next subsection: (i) linear regularized models (Lasso and $\ell_{2,1}$) (ii) neural network (iii) long-short term memory neural network (LSTM), and (iv) convolutional neural network (CNN).

C. Multi-Task Regularized Linear Models

Linear models are interpretable and widely-used in solving regression and classification problems. We apply two regularized MTL linear regression and classification models (MTL Lasso and MTL $\ell_{2,1}$). For MTL ℓ_1-norm, known as Lasso regularization, introduces the sparsity into the models and reduces complexity of learning. The MTL $\ell_{2,1}$-norm regularization selects features jointly so that all tasks have same sparsity on features. These two models select features automatically and we can interpret the contributing features to each wellbeing label.

The objective function for MTL Linear model:

$$\hat{\theta} = argmin_\theta \sum_{i=1}^{n} loss\{\theta_i^T X_i, Y_i\} + \{\lambda_1 ||\theta||_1, or\ \lambda_{2,1}||\theta||_{2,1}\}$$

The loss function for linear regression is: $||\theta_i^T X_i - Y_i||_2^2$; The loss function for logistic regression(classification) is: $\sum_{j=1}^{m_i} log(1 + exp(-Y_{i,j}(\theta_j^T X_{i,j} + c_i)))$ where X_i represents the input matrix of the i-th task, and Y_i is the label of the corresponding task. θ is the matrix of weights of the model. λ_1 and $\lambda_{2,1}$ are Lagrange multipliers for each of algorithms as the part of regularized terms.

For the Lasso model, the regularization parameter λ_1 of the norm term controls the sparsity of weights for the single task; whereas in the $\ell_{2,1}$-norm form, $\lambda_{2,1}$ controls the sparsity for all tasks together. We tuned the parameters via grid search (λ_1:0.005, $\lambda_{2,1}$: 200). Linear models were implemented via MALSAR [16].

D. Multi-Task Neural Network

Building a neural network for multitasking is building a network structure that can handle multiple inputs and multiple outputs, with hidden layers that are shared by all tasks (Fig.1). We used the shared hidden layers to summarize and extract participants both common and different characteristics. For example, mood may be influenced by sleep behavior on previous nights for almost everyone but the frequency of phone usage might only influence specific participants' mood.

In order to avoid overfitting, a dropout method was used with a factor of 0.5 after grid searching.

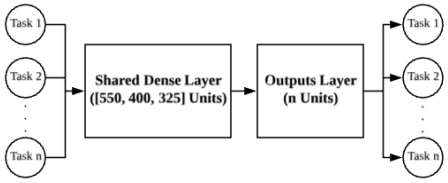

Fig. 1. Structure of Multi-Task Neural Network

E. Recurrent Neural Network

Long Short-Term Memory (LSTM) network structures is an extended structure of RNN that can learn long-term dependencies [17] and has been used in time-series problem such as natural language processing [18].

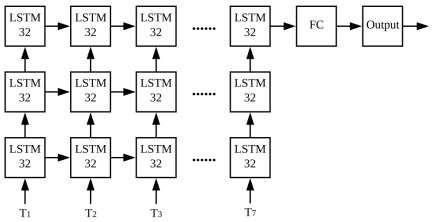

Fig. 2. Structure of long-short term memory neural network. T represents time steps. FC is the fully connected layer.

We used a multi-layer LSTM for sequential learning using each participants previous 7 days of data, with one day as per time step (Fig. 2). After grid searching though cross-validation, we adopted the following structure and parameters: 3 layers of LSTM with 32 recurrent units are connected, incorporating 0.4 recurrent dropout and 0.25 dropout rates in each layer. This is followed by a fully-connected layer with 250 hidden units with a dropout rate 0.5. The Adam optimizer with learning rate 0.005 was also adopted.

We extended this LSTM structure for MTL and followed the multi-task LSTM method used by H. Suresh *et al* [15]. There are two approaches to designing the multi-task neural network models: (i) the output is generated directly from a large shared full connection layer, or (ii) there are unique hidden layers for each output. For our prediction tasks, we chose the first approach because in our experiments comparing these two approaches on our dataset, we found that (i) due to the constraints of the data volume, overfitting problem of unique layers model led to similar performance of two, and (ii) the training time of the second model was significantly longer than that of the first model.

F. Convolutional Neural Network (CNN)

CNNs are actively used in deep learning. We use convolution kernels not only to process features in the input matrix but also to extract relationships among them across the time steps. After designing and experiments, we chose an MTL CNN structure as shown in Figure 3. In this structure, the input matrix is composed of the observations(participants) $\times$ features with time steps and outputs will be the prediction

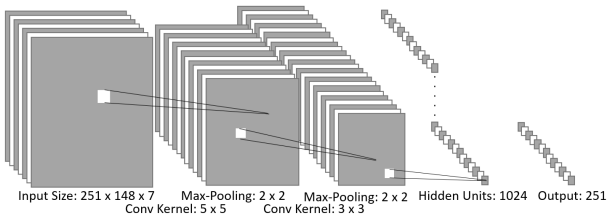

Fig. 3. The structure of Convolutional Neural Network

vector with elements corresponding to each participant. Theoretically, this structure takes the advantages of both MTL and time series learning. With dropout probability 0.25 in the dense layer and 0.01 weighted ℓ_2 penalty terms in each convolution layer, we also alleviated the over-fitting problem.

A non-MTL CNN model which inputs the data of time steps $\times$ features was build as a comparison. We used a 7×7 kernel conv layers with a 3×3 max-pooling layer, a 5×5 kernel conv layers with a 2×2 max-pooling layer followed by a 512 hidden units dense layer, and the output size was 1. We also applied weighted ℓ_2 penalty terms by 0.01 and 0.25 dropout.

G. Imputation

The dataset has missing data for multiple reasons (e.g., sensor broke, participant did not complete a diary or forgot to wear sensors). The previous work has demonstrated that an auto-encoder approach can result in better performance than mean or median imputation especially when a larger portion of data are missing [19]. An auto-encoder is an artificial neural network to extract a representation in an unsupervised method [20]. One application of an auto-encoder is the denoising auto-encoder (DAE), which can reconstruct data corrupted by data-missing noise. In this project, we implement the DAE in Keras [14] to process the missing values.

III. Experiment

Our tasks are formulated in two ways for evaluation. As a regression, the problem is to predict next day's mood, health and stress wellbeing scores, each in the range of 0-100. As a classification, the problem is to predict for each next day's wellbeing score whether it will be in the range high, mid, or low (defined here as 100-67, 66-34, or 33-0). In both problems, the system can use the data up to and including the current day in formulating its prediction for next day's label. We conducted several different comparisons: (i) We compared the performances of using different modalities of features to train the model; (ii) We compared the number of previous N days of data (current day, previous 3, 5, 7 days) to predict next day's mood, health and stress and (iii) We compared predicting N days into the future.

We shuffled the dataset 5 times for each person in all train/validation/test datasets, and applied a cross validation method in parameters searching and training the models. We evaluated mean absolute errors for the regression models and F1 scores for classification tasks (i.e., train/validation/test: 60%/20%/20%). Principal component analysis (PCA) was applied with 0.99 of explained variance for dimensionality reduction in each fold of cross validation. After PCA, there are 148 features left per day across 251 participants. In addition, we adopted focal loss [21] as the objective function in the classification tasks to mitigate the unbalanced sample size in the 3 classes, as the low:mid:high ratios were 3:9:8, 1:3:3, and 5:9:6, for mood, health, and stress respectively. The Adam optimizer [22] was used in training the neural networks, with a learning rate of 0.005 and 0.9, 0.999 for β_1 and β_2.

TABLE I
Prediction Performance with Different Algorithms
(MTL:Persons as tasks, Regression[Mean Absolute Value] & Classification[F1 Score])

Algorithms		Mood±(SD)	Health±(SD)	Stress±(SD)
Linear Models (Regression)	MTL Lasso	13.7 (0.3)	13.5 (0.3)	15.4 (0.3)
	MTL $\ell_{2,1}$	12.8 (0.3)	11.9 (0.2)	13.7(0.4)
Neural Networks (Regression)	MTL NN	12.8 (0.2)	12.8 (0.6)	13.7 (0.3)
	LSTM	14.5 (0.3)	12.4 (0.5)	15.3 (0.5)
	MTL LSTM	13.2 (0.1)	12.4 (0.4)	14.9 (0.4)
	CNN	18.4 (0.5)	17.6 (0.7)	19.1 (0.5)
	MTL-CNN	13.9 (0.7)	13.0 (0.6)	14.5 (0.7)
Logistic Models (Classification)	MTL Lasso	0.63 (0.01)	0.67 (0.01)	0.59 (0.01)
	MTL $\ell_{2,1}$	0.65 (0.01)	0.68 (0.01)	0.61(0.02)
Neural Networks (Classification)	MTL NN	0.71 (0.01)	0.74 (0.01)	0.66 (0.02)
	LSTM	0.66 (0.02)	0.73 (0.01)	0.63 (0.01)
	MTL LSTM	0.69 (0.02)	0.74 (0.01)	0.65 (0.01)
	CNN	0.51 (0.01)	0.53 (0.01)	0.50 (0.01)
	MTL-CNN	0.68 (0.01)	0.72 (0.01)	0.67 (0.01)

IV. Results & Discussion

The next-day prediction results from regression and 3-class classification using the past-7-day data are shown in Table I. In the linear regularized models, the $\ell_{2,1}$ method has better prediction performance and shows statistically significantly lower mean absolute errors compared with the Lasso algorithm (paired t-test; $p<0.05$). In the neural network models, the MTL-NN showed a significantly better performance for mood and stress prediction in both regression and classification task (ANOVA, Tukey; $p<0.05$); whereas MTL-LSTM produced significantly better health prediction results ($p < 0.05$). Overall, $\ell_{2,1}$ linear model performed the best for the regression task and MTL-NN had a better performance for mood and stress prediction for the classification task, while MTL-LSTM performed the best for health prediction. Limited by the size of data set, LSTM and MTL-LSTM structure have the over-fitting problem; whereas linear model and MTL-NN can achieve better results by regularization in the labels of mood and stress.

Our results showed that MTL performed significantly better in our dataset, especially with persons as tasks rather than groups of people as tasks. Our MTL $\ell_{2,1}$ regression models - groups of people as tasks (clustering based on personalities surveys and genders with highest Silhouette score, 251 participants were divided into 21 groups) showed 14.7, 14.4 and 16.2 for mood, health and stress, which are less effective than MTL $\ell_{2,1}$ - persons as tasks($p < 0.05$). And our performances in regression tasks also show improvements compared to the prior persons-as-tasks work whose average MAE values of mood, health and stress were 13.0, 12.9 and 14.1 [8].

We analyzed when our models showed high and low errors. In regression, we observed larger errors when the labels in the previous days had larger fluctuations. In the confusion matrices of classification tasks, we found that our models have higher performance at predicting mid & high mood, high health, and low stress classes by 10-19% compared to other ranges in these labels. We also examined top/mid/bottom 33% label split models and found 10% lower F1 scores compared to the 100-67/66-34/33-0 data split pattern. By examining the confusion matrix, we found that the main reason for this change was the prediction performance of the mid labels was lower.

In the classification task, high accuracy (100%) was also achieved in only 18-39 participants. For the health label, the prediction accuracy on 39 of the participants were 100%, while that on one participant was only 16%. We found that the participants on whom the methods had better performance had flat labels in training (mean: 68.4; SD:7.7) and testing data (mean:68.3; SD:6.3), while the participants on whom our methods often made errors had widely fluctuating labels in training (mean:45.8; SD:20.8) and testing (mean:61.9; SD:16.2) data. We also found that there are statistical differences in some features between participants on whom we had good prediction accuracy and those on whom we had bad accuracy. For example, in the mood prediction, there were statistically significant differences between good and poor accuracy in the median EDA signal of participants and the use of mobile phone screen between 3-10 am ($p < 0.05$, respectively). Our results also showed that the health prediction models performed better than the stress models. We compared the mean and variance of the two labels and found that the health labels (mean:64.3, SD:24.9) have higher mean value and lower SD than the stress labels (mean:52.6, SD:26.2) and models are able to predict better with low SD health labels.

We found that the combination of all modalities: surveys, wearable, mobile phone and weather performed the best but the combination of wearable, weather and mobile phone features or the combination of wearable and mobile phone performed better than wearable only or phone only. In the comparison of the prediction performance using the previous 1, 3, 5 or 7 days of data, mood and health regression and classification models performed the best with 7 days of data. Stress models performed the best with 5 and 7 days of data (one-way ANOVA, tukey test, $p < 0.05$). This suggests that the length of the data to approximately 7 days improves the prediction performance; this result was consistent with a previous study that showed that human wellbeing is affected by week-long weather and behaviors [23]. In addition to predicting next day's wellbeing, we predicted mood, health and stress scores for the next 3, 5 and 7 days using the best performing regression and classification algorithms, MTL linear model and MTL NN. In the regression task, we found that the prediction of next day's labels generally had significantly smaller MAE values (0.5-0.8) than the predictions of further into the future labels($p < 0.05$).

One of the advantages of the linear model is that we can observe the weights of the training model and then analyze the most important features of the model. MTL $\ell_{2,1}$ linear model performed the best in the regression task. In the health model, weather features (such as air quality of day, visibility, and air pressure) and no scheduled activities in the morning showed higher weights. In the stress model, both weather factors and academic activity showed higher contributions. Our finding matched prior results that having fewer academic activities is associated with less stress for students [24]. Our findings are also consistent with prior work showing weather impacts human mood [26], health, and stress [25].

V. Conclusion

In this paper, we compared the performance of mood, health and stress prediction models using several MTL and deep learning algorithms. Our results showed that MTL-linear models, and MTL-NN and MTL-LSTM performed the best for regression and 3-class classification. Our models showed that weather features and features about first schedules in the morning and academic activities contributed highly to well-being labels. Our analysis also showed that high fluctuations in the previous days of wellbeing labels were associated with larger errors in the prediction models. As future work, we will investigate other model structures such as a CNN-LSTM model and validate our models with data from other student and non-student populations.

References

[1] H.L. Dunn "High level wellness." Oxford, England: R. W. Beaty, 1973.
[2] R. Dodge *et al.* "The Challenge of Defining Wellbeing," International Journal of Wellbeing, 2(3), 222-235.
[3] S. Stewart-Brown, "Emotional wellbeing and its relation to health. Physical disease may well result from emotional distress," BMJ. 1998
[4] R. Cai *et al.* "Correlational Analyses among Personality Traits, Emotional Responses and Behavioral States Using Physiological Data from Wearable Sensors," TELEMED 2018
[5] M. Matthews *et al.* "Tracking Mental Well-Being: Balancing Rich Sensing and Patient Needs," Computer,vol. 47, no. 4, pp. 36-43, 2014.
[6] 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 2018;20(6):e210.
[7] S.A. Taylor *et al.* "Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health," in IEEE TAC 2017.
[8] N.Jaques *et al.* "Predicting Tomorrows Mood, Health, and Stress Level using Personalized Multitask Learning and Domain Adaptation," AffComp at IJCAI, 2017.
[9] T. Umematsu *et al.* "Improving Stress Forecasting using LSTM Neural Networks." IEEE EMBC, 2018.
[10] O.P. John *et al.* "The Big Five Trait taxonomy: History, measurement, and theoretical perspectives." Handbook of personality
[11] The Dark Sky Company LLC. Dark Sky Forecast API, 2016. URL https://developer.forecast.io/.
[12] R. Caruana "Machine Learning", (1997) Kluwer Academic Publishers.
[13] A. Liu *et al.* "Hierarchical Clustering Multi-Task Learning for Joint Human Action Grouping and Recognition," in IEEE PAMI 2017.
[14] F. Chollet. "Keras." https://github.com/ fchollet/keras, 2015.
[15] H. Suresh *et al.* "Learning Tasks for Multitask Learning: Heterogenous Patient Populations in the ICU," KDD 2018
[16] J. Zhou *et al.* "Malsar: Multi-task learning via structural regularization," Arizona State University 21
[17] S. Hochreiter *et al.* "Malsar: Long Short-Term Memory," Neural Computation 1997 Vol. 9, 1735-1780
[18] K. Cho *et al.* "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," EMNLP 2014
[19] N. Jaques *et al.* "Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction," IEEE ACII 2017
[20] C.Y. Liou *et al.* "Autoencoder for words," Neurocomputing 2014
[21] T. Lin *et al.* "Focal Loss for Dense Object Detection," ICCV 2017.
[22] D.P. Kingma, "Adam:A Method For Stochastic Optimization" ICLR 2015
[23] F. Bentley *et al.* "Health Mashups: Presenting Statistical Patterns between Wellbeing Data and Context in Natural Language to Promote Behavior Change," ACM TOCHI 2013.
[24] K.J. Reddy *et al.* "Academic Stress and its Sources Among University Students," Biomed Pharmacol J 2018;11(1).
[25] F.G. Sulman "The impact of weather on human health," Reviews on Environmental Health, 1984
[26] T. Klimstra *et al.* "Come rain or come shine: individual differences in how weather affects mood," Emotion, 2011.