

Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health

Sara Taylor*, *Student Member, IEEE*, Natasha Jaques*, *Student Member, IEEE*,
Ehimwenma Nosakhare, *Student Member, IEEE*, Akane Sano, *Member, IEEE*,
and Rosalind Picard, *Fellow, IEEE*

Abstract—While accurately predicting mood and wellbeing could have a number of important clinical benefits, traditional machine learning (ML) methods frequently yield low performance in this domain. We posit that this is because a one-size-fits-all machine learning model is inherently ill-suited to predicting outcomes like mood and stress, which vary greatly due to individual differences. Therefore, we employ Multitask Learning (MTL) techniques to train personalized ML models which are customized to the needs of each individual, but still leverage data from across the population. Three formulations of MTL are compared: i) MTL deep neural networks, which share several hidden layers but have final layers unique to each task; ii) Multi-task Multi-Kernel learning, which feeds information across tasks through kernel weights on feature types; and iii) a Hierarchical Bayesian model in which tasks share a common Dirichlet Process prior. We offer the code for this work in open source. These techniques are investigated in the context of predicting future mood, stress, and health using data collected from surveys, wearable sensors, smartphone logs, and the weather. Empirical results demonstrate that using MTL to account for individual differences provides large performance improvements over traditional machine learning methods and provides personalized, actionable insights.

Index Terms—Mood Prediction, Multitask learning, Deep Neural Networks, Multi-Kernel SVM, Hierarchical Bayesian Model

1 INTRODUCTION

PERCEIVED wellbeing, as measured by self-reported health, stress, and happiness, has a number of important clinical health consequences. Self-reported happiness is not only indicative of scores on clinical depression measures [1], but happiness is so strongly associated with greater longevity that the effect size is comparable to that of cigarette smoking [2]. Stress increases susceptibility to infection and illness [3]. Finally, self-reported health is so strongly related to actual health and all-cause mortality [4], that in a 29-year-study it was found to be the single most predictive measure of mortality, above even more objective health measures such as blood pressure readings [5].

Clearly, the ability to model and predict subjective mood and wellbeing could be immensely beneficial, especially if such predictions could be made using data collected in an unobtrusive and privacy-sensitive way, perhaps using wearable sensors and smartphones. Such a model could open up a range of beneficial applications which passively monitor users' data and make predictions about their mental and physical wellbeing. This could not only aid in the management, treatment, and prevention of both mental illness and disease, but the predictions could be useful to any person who might want a forecast of their future mood, stress, or health in order to make adjustments to their routine to attempt to improve it. For example, if the model predicts

that I will be extremely stressed tomorrow, I might want to choose a different day to agree to review that extra paper.

Unfortunately, modeling wellbeing and mood is an incredibly difficult task, and a highly accurate, robust system has yet to be developed. Historically, classification accuracies have ranged from 55-80% (e.g., [6]–[8]), even with sophisticated models or multi-modal data. In this paper, we use a challenging dataset where accuracies from prior efforts to recognize wellbeing and mood ranged from 56-74% [9], [10]. Across many mood detection systems, performance remains low despite researchers' considerable efforts to develop better models and extract meaningful features from a diverse array of data sources.

We hypothesize that these models suffer from a common problem: the inability to account for individual differences. What puts one person in a good mood does not apply to everyone else. For instance, the stress reaction experienced by an introvert during a loud, crowded party might be very different for an extrovert [11]. Individual differences in personality can strongly affect mood and vulnerability to mental health issues such as depression [12]. There are even individual differences in how people's moods are affected by the weather [13]. Thus, a generic, omnibus machine learning model trained to predict mood is inherently limited in the performance it can obtain.

We show that accounting for interindividual variability via MTL can dramatically improve the prediction of these wellbeing states: mood, stress, and health. MTL is a type of transfer learning, in which models are learned simultaneously for several related tasks, but share information through similarity constraints [14]. We show that MTL can allow each person to have a model tailored specifically for them, which still learns from all available data. Therefore,

- S. Taylor, N. Jaques, A. Sano and R. Picard are with the Program of Media Arts and Sciences and the MIT Media Lab.
E-mail: {sataylor, jaquesn, akane, picard}@media.mit.edu
- E. Nosakhare is with the Department of Electrical Engineering and Computer Science and the MIT Media Lab. E-mail: ehinosa@mit.edu

*Both authors contributed equally to this work

Manuscript received June 28, 2017; revised Nov 2, 2017; Accepted Dec 2 2017.

the approach remains feasible even if there is insufficient data to train an individual machine learning model for each person. By adapting existing MTL methods to account for individual differences in the relationship between behavior and wellbeing, we are able to obtain state-of-the-art performance on the dataset under investigation (78-82% accuracy), significantly improving on prior published results.

In addition to showing the benefits of personalization, we undertake a more challenging task than is typically attempted when modeling mood. While most prior work has focused on *detecting* current mood state, we test the ability to *predict* mood and wellbeing tomorrow night (at least 20 hours in the future), using only data from today. Specifically, assume x_t represents all the smartphone, wearable sensor, and weather data collected about a person on day t (from 12:00am to 11:59pm). Let y_t be the person's self-reported mood, stress, and health in the evening of day t (reported after 8pm). Previous work has focused on learning to model $p(y_t|x_t)$; that is, the probability of the person's current mood given the current data, which we refer to as mood *detection*. In contrast, we learn $p(y_{t+1}|x_t)$, the probability of the person's mood tomorrow given today's data. This type of prediction could be considered a type of mood *forecasting*, providing an estimate of a person's future wellbeing which could potentially allow them to better prepare for it — just as a weather forecast gives one a chance to take an umbrella rather than being left to be soaked by the rain.

Typical forecasting models make use of a history of prior labels to make predictions; i.e. such models learn the function $p(y_{t+1}|y_t, y_{t-1}, \dots, y_1, x_t, x_{t-1}, \dots, x_1)$. Using such a model for mood forecasting is less than desirable, since it implies that a person must input their mood every day in order to obtain predictions about tomorrow. In contrast, we do not use any prior labels. Instead, we learn $p(y_{t+1}|x_t)$, allowing us to predict an individual's mood without ever requiring them to manually input a mood rating.

Our work makes the following contributions to the affective computing literature. We predict future wellbeing without requiring a history of collected wellbeing labels for each person. Our data are gathered in the “wild” as participants go about their daily lives, using surveys, wearable sensors, weather monitoring, and smartphones, and thus are relevant to use in a real-world wellbeing prediction system. We provide insights into the relationship between the collected data and mood and wellbeing, through an implicit soft clustering of users provided by the Bayesian model, and a learned weighting of input sources. Finally, we demonstrate the ability of MTL to train personalized models that can account for individual differences, and provide the developed code for the MTL models in open source. The insight that personalization through MTL can significantly improve mood prediction performance could be a valuable step towards developing a practical, deployable mood prediction system.

2 RELATED WORK

The idea of estimating mood, stress, and mental health indicators using unobtrusive data collected from smartphones and wearables has been garnering increasing interest. For example, Bogomolov and colleagues [15] use smartphone

data combined with information about participants' personalities and the weather to detect stress with 72% accuracy. Other researchers have investigated using smartphone monitoring to detect depressive and manic states in bipolar disorder, attaining accuracy of 76% [8]. Detecting workplace stress is another growing body of research [16].

The insight that an impersonal, generic affect classifier cannot account for individual differences has been arrived at by several researchers. In estimating workplace stress, Koldijk and colleagues found that adding the participant ID as a feature to their model could improve accuracy in classifying mental effort [17]. Similarly, Canzian and colleagues found that training a generic SVM to classify depressive mood from location data and surveys resulted in sensitivity and specificity values of 0.74 and 0.78, respectively. By training an independent SVM for each person, the authors obtained values of 0.71 and 0.87 [18].

Finally, a detailed study reported that an omnibus model trained to detect all people's mood based on smartphone communication and usage resulted in a prediction accuracy of 66% [7]. However, if two months of labeled data were collected for each person, then individual, independent personalized models could be trained to achieve 93% accuracy in mood classification! Since obtaining two months of training data per person can be considered somewhat unrealistic, the authors investigated methods for training a hybrid model that weights personalized examples more heavily, which can be used when there are fewer labeled training examples per person. In contrast with this work we focus on methods for making reasonable personalized predictions even in the absence of any labeled training data for a new person.

As mentioned above, almost all prior work of which we are aware has focused on mood *detection*, rather than true prediction; that is, learning $p(y_t|x_t)$, where the model label y_t and data x_t are both collected on day t . A recent paper published in April 2017 claims to be the first work to forecast future mood [19]. This work uses a Recurrent Neural Network (RNN) to predict mood given two weeks of mood history reported every day, learning the function $p(y_{t+1}|y_t, y_{t-1}, \dots, y_1, x_t, x_{t-1}, \dots, x_1)$. Using a large-scale dataset of 2,382 people, the authors achieved an AUC score of 0.886 in forecasting severely depressed mood. While a notable contribution, the drawback to this approach is that it requires a person to diligently input their mood every day. If one day is missed, a prediction cannot be made for the next two weeks. Further, the results reveal that past mood features are many times more effective at predicting future mood than any of the other data collected. Thus, using a mood history to predict future mood is a significantly easier problem. In contrast, we are able to predict tomorrow's wellbeing given a rich set of data from today ($p(y_{t+1}|x_t)$), obtaining accurate predictions about an individual's future mood through personalization, without requiring them to manually input self-reported labels.

2.1 Multitask Learning

MTL is a type of transfer learning, in which models are learned simultaneously for several tasks but share information through similarity constraints. Originally proposed as a

way to induce efficient internal representations within neural networks (NNs) [14], MTL can be used across a variety of models. It can be considered a form of regularization, and can improve generalization performance [14] as long as tasks are sufficiently related [20]. Because MTL is beneficial when training data is scarce and noisy, it is well-suited to the messy, real-world problem of predicting mood.

Since Caruana's original work, a variety of NN MTL methods have been explored. For example, face detection accuracy for a deep convolutional network can be improved by sharing layers with networks trained on similar tasks, like face pose estimation and facial landmark localization [21]. Multitasking has also been used successfully to train NNs with very little data; by using the same network to predict traffic flow in the past, present, and future, Jin and Sun were able to improve prediction accuracies using only 2112 samples of traffic data [22].

Hierarchical Bayesian learning is a popular approach to MTL; Baxter and colleagues provide a detailed overview [23]. The general approach is exemplified by an algorithm like *Transfer-Aware Naïve Bayes* [20]: each task's model's parameters are drawn from a common prior distribution, thus imposing a similarity constraint. The model can update the parameters of the prior as it learns — for example, by decreasing the variance if the tasks are very similar. Bayesian inference techniques have been applied to a number of MTL scenarios. For example, MTL has been applied to a reinforcement learning problem in which each task is an environment an agent must explore, and the Markov Decision Process (MDP) learned for previous environments is treated as a strong prior on the model for a new environment [24].

MTL has also been explored within the Affective Computing community. The idea of treating predicting the affect of a single person as a task was introduced in conjunction with Multi-Task Multi-Kernel Learning (MTMKL) [25] using the DEAP dataset [26]. MTKML is an MTL method specifically designed for Affective Computing applications which need to combine data from multiple disparate sources, or modalities [25]. A kernel function is computed using the features from each modality, and these are combined in a weighted sum. MTL is applied by learning separate kernel weights for each task, while constraining all tasks' weights to be similar. Thus, information is shared across tasks through the kernel weights on the modalities [25]. While treating modeling different people as related tasks in MTKML allows for personalization, it does not allow the model to generalize to a new person. In contrast, we first cluster people based on personality and treat predicting the wellbeing of a cluster as a task, allowing us to generalize to new users who have not input wellbeing labels by placing them into the appropriate cluster.

MTL can also be applied to Affective Computing by treating outcomes like arousal and valence as the related tasks in the model. This method was used in our prior work, which applied MTKML to the dataset under investigation in this paper by treating the classification of happiness, stress, health, alertness, and energy as related tasks [10]. Similarly, Xia and colleagues improved the performance of a deep belief network by training it to simultaneously recognize both valence and arousal from speech [27]. In another study of speech emotion recognition, the authors found that treating

different corpora, domains, or genders as related tasks in an MTL framework offered performance benefits over learning a single model over all of the domains, or learning a separate model for each domain [28].

3 METHODS

In what follows we describe several techniques for using MTL to account for interindividual differences in the relationship between behavior, physiology, and resulting mood and wellbeing. Each of the models can adapt to the specific characteristics of each person, while still sharing information across people through a) shared layers of a deep neural network (Section 3.1); b) a similarity constraint on each task's classifier's weights (Section 3.2); or c) a common prior shared across tasks (Section 3.3)

The most intuitive way to use MTL to customize a model for each person is to treat predicting the wellbeing of a single person as a single task. However, this approach may become untenable if there are few samples per person. Since it requires that each person have a unique, fully trained, task-specific model, each person would need to provide a sufficient amount of labeled data. This may constitute a disincentivizing burden on potential users. More importantly, such a model *cannot generalize to new people or make accurate predictions about new users*.

Therefore, we begin by clustering users based on their personality and gender, and treat predicting mood for a given cluster as one prediction task. In this way, we can easily make predictions for a new user without requiring them to input their mood on a daily basis; we simply use their personality and gender to assign them to the appropriate cluster. In this study, personality is computed using the Big Five trait taxonomy [29], via a questionnaire that takes approximately 10 minutes to complete. We apply K-means clustering to participants' Big Five scores and gender, and assess cluster quality in an unsupervised way using *silhouette score*, which is evaluated based on the intra-cluster and nearest-cluster distance for each sample [30]. The number of clusters which produced the highest silhouette score was $K = 37$. Using a large number of clusters allows us to create fine-grained, highly customized models that make mood predictions for extremely specific types of people.

3.1 Neural Networks

To build a MTL neural network (NN), we begin with several initial hidden layers that are shared among all the tasks. These layers then connect to smaller, task-specific layers, which are unique to each cluster. Figure 1 shows a simplified version of this architecture. In reality, the network can have many shared and task-specific layers.

The intuition behind this design is that the shared layers will learn to extract information that is useful for summarizing relevant characteristics of any person's day into an efficient, generalizable embedding. The final, task-specific layers are then expected to learn how to map this embedding to a prediction customized for each cluster. For example, if the shared layers learn to condense all of the relevant smartphone app data about phone calls and texting into an aggregate measure of social support, the task-specific

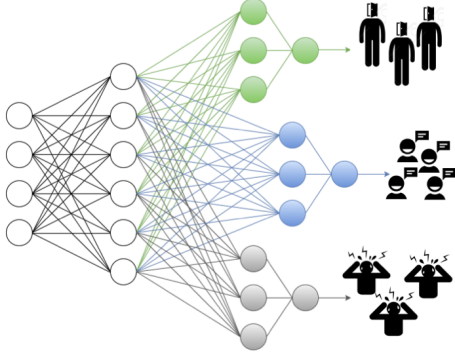


Fig. 1: A simplified version of the MTL-NN architecture. Clusters of related people receive specialized predictions from a portion of the network trained with only their data. Shared initial layers extract features relevant to all clusters.

layers can then learn a unique weighting of this measure for each cluster. Perhaps a cluster containing participants with high extroversion scores will be more strongly affected by a lack of social support than another cluster.

To train the network, we must slightly modify the typical Stochastic Gradient Descent (SGD) algorithm. Rather than randomly selecting a mini-batch of N training samples from any of the available data, each mini-batch must only contain data from a single randomly selected task or cluster. The mini-batch is then used to predict label values y'_i , based on forward propagation through the shared weights and the appropriate cluster-specific weights. The ground-truth target labels y_i are used to compute the error with respect to that batch using the typical cross-entropy loss function:

$$L_H(\mathbf{Y}, \mathbf{Y}') = - \sum_{i=1}^N [y_i \log y'_i + (1 - y_i) \log(1 - y'_i)]$$

A gradient step based on the loss is then used to update both the cluster-specific weights, as well as to adjust the weights within the shared layers. By continuing to randomly sample a new cluster and update both the cluster-specific and shared weights, the network will eventually learn a shared representation relevant to all clusters.

While deep learning is a powerful branch of ML, when training on small datasets such as the one under discussion in this paper it is important to heavily regularize the network to avoid overfitting. Although MTL itself is a strong form of regularization, we implement several other techniques to ensure generalizable predictions. As is common, we include the following L2 regularization term in the loss function: $-\beta \|\mathbf{W}\|_2^2$, where \mathbf{W} are the weights of the network. We also train the network to simultaneously predict all three wellbeing labels to further improve the generalizability of the embedding. Finally, we implement dropout, a popular approach to NN regularization in which some portion of the network's weights are randomly "dropped out" (set to 0) during training. This forces the network to learn redundant representations and is statistically very powerful. Using a dropout factor of 0.5 (meaning there is a 50% chance a given weight will be dropped during training) on a NN with n nodes is equivalent to training 2^n NNs which all share parameters [31]. This is easy to verify; consider a binary variable that represents whether or not a

node is dropped out on a given training iteration. Since there are n nodes, there are 2^n possible combinations of these binary variables. Moreover, each of these sub-networks are trained on different, random mini-batches of data, and this bagging effect further enhances generalization performance.

3.2 Multi-Task Multi-Kernel Learning

As introduced in Section 2, the MTMKL algorithm developed by Kandemir et. al. [25] is a MTL technique designed for the problem of classifying several related emotions (tasks) based on multiple data modalities [25]. MTMKL is a modified version of Multi-Kernel Learning (MKL) in which tasks share information through kernel weights on the modalities. Here, we consider the problem of using MTMKL to build a personalized model that can account for individual differences. Therefore, we treat each task as predicting the wellbeing for one cluster; that is, a group of highly similar people which share the same gender or personality traits.

MTMKL uses a least-squares support vector machine (LSSVM) for each task-specific model. Unlike the canonical SVM, the LSSVM uses a quadratic error on the "slack" variables instead of an L1 error. As a result, the LSSVM can be learned by solving a series of linear equations in contrast to using quadratic programming to learn an SVM model. The LSSVM has the added benefit that when only a single label is present in the training data, its predictions will default to predict only that label.

The LSSVM can be learned by solving the following optimization problem, in which N is the total number of samples, x_i is the i th feature vector, y_i is the i th label, $k(x_i, x_j)$ is a kernel function, and α is the set of dual coefficients as in a conventional SVM:

$$\begin{aligned} \underset{\alpha}{\text{maximize}} \quad & -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j k(x_i, x_j) \\ & - \frac{1}{2C} \sum_{i=1}^N \alpha_i^2 + \sum_{i=1}^N \alpha_i \\ \text{subject to} \quad & \sum_{i=1}^N \alpha_i y_i = 0 \end{aligned}$$

In MKL, we build on the LSSVM by adjusting the kernel function. In particular, we use a kernel function to compute the similarity between feature vectors for each modality m , and the kernels are combined using a weighted sum. The weights depend on the usefulness of each modality in prediction. That is, more useful modalities will have larger kernel weights so that differences in that data modality are more helpful in prediction.

Concretely, we assign a kernel k_m to the features in modality m , as in typical MKL. We restrict the model space by using the same kernel function (e.g., an RBF kernel) for each modality. The modality kernels are combined into a single kernel, k_η , in a convex combination parameterized by the kernel weighting vector, η . Let $\mathbf{x}_i^{(m)}$ be the i th feature vector that contains only the features belonging to modality m , and M be the total number of modalities. Then k_η is defined as follows:

$$k_{\eta}(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\eta}) = \sum_{m=1}^M \eta_m k_m(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)})$$

such that $\eta_m > 0, m = 1, \dots, M$ and $\sum_{m=1}^M \eta_m = 1$. Thus the LSSVM-based MKL model can be learned using the same optimization as the LSSVM with the additional constraint of the convex combination of kernel weights $\boldsymbol{\eta}$.

When multiple tasks are learned at the same time in MTMKL, each task t has its own vector of kernel weights, $\boldsymbol{\eta}^{(t)}$, which are regularized globally by a function which penalizes divergence from the weights of the other tasks. This allows information about the kernel weights to be shared between tasks so that each task benefits from the data of other tasks. In particular, if the model is highly regularized, then the kernel weight on the m th modality (i.e., $\eta_m^{(t)}$) will be very similar across all tasks t . As such, each task will treat the modalities as having similar importance. Note that even though the kernel weights might be highly regularized, the task-specific models can still learn a diverse set of decision boundaries within the same kernel space.

The optimal $\boldsymbol{\eta}^{(t)}$ for all tasks $t = 1, \dots, T$ can be learned by solving a min-max optimization similar to the LSSVM-based MKL model, but with the addition of the regularization function, $\Omega(\{\boldsymbol{\eta}^{(t)}\}_{t=1}^T)$. A weight ν placed on the regularization function $\Omega(\cdot)$ controls the importance of the divergence. When $\nu = 0$ the tasks are treated independently, and as ν increases, the task weights are increasingly restricted to be similar.

For simplicity of notation we denote the objective function for a single task's LSSVM-based MKL model as follows:

$$J^{(t)}(\alpha^{(t)}, \boldsymbol{\eta}^{(t)}) = -\frac{1}{2} \sum_{i=1}^{N^{(t)}} \sum_{j=1}^{N^{(t)}} \alpha_i^{(t)} \alpha_j^{(t)} y_i y_j k_{\boldsymbol{\eta}^{(t)}}(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{2C} \sum_{i=1}^{N^{(t)}} \alpha_i^2 + \sum_{i=1}^{N^{(t)}} \alpha_i$$

where the superscript (t) denotes the parameters or functions specific to task t .

Thus, all of the parameters of the LSSVM-based MTMKL model can be learned by solving the following min-max optimization problem:

$$\begin{aligned} & \underset{\{\boldsymbol{\eta}^{(t)}\}_{t=1}^T}{\text{minimize}} \quad \underset{\{\alpha^{(t)}\}_{t=1}^T}{\text{maximize}} \quad \nu \Omega(\{\boldsymbol{\eta}^{(t)}\}_{t=1}^T) + \sum_{t=1}^T J^{(t)}(\alpha^{(t)}, \boldsymbol{\eta}^{(t)}) \\ & \text{subject to} \quad \sum_{i=1}^N \alpha_i y_i = 0 \\ & \quad \sum_{m=1}^M \eta_m^{(t)} = 1, t = 1, \dots, T \\ & \quad \eta_m^{(t)} \geq 0, \forall m, \forall t \end{aligned}$$

The iterative gradient descent method proposed by Kandemir et. al [25] is used to train the model given an initial set of model parameters. The method alternatively (1) solves a LSSVM for each task given $\boldsymbol{\eta}^{(t)}$ and (2) updates $\boldsymbol{\eta}$ in the direction of negative gradient of the joint objective function (see Algorithm 1).

Let the joint objective function be $O_{\boldsymbol{\eta}}$. We write the gradient as follows:

$$\begin{aligned} \frac{\partial O_{\boldsymbol{\eta}}}{\partial \eta_m^{(t)}} &= \nu \frac{\partial}{\partial \eta_m^{(t)}} \Omega(\{\boldsymbol{\eta}^{(t)}\}_{t=1}^T) \\ &\quad - \frac{1}{2} \sum_{i=1}^{N^{(t)}} \sum_{j=1}^{N^{(t)}} \alpha_i^{(t)} \alpha_j^{(t)} y_i^{(t)} y_j^{(t)} k_m(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)}) \end{aligned}$$

Algorithm 1 MTMKL Algorithm

- 1: Initialize $\boldsymbol{\eta}^{(t)}$ as $(1/T, \dots, 1/T), \forall t$
 - 2: **while** not converged **do**
 - 3: Solve each LSSVM-based MKL model using $\boldsymbol{\eta}^{(t)}, \forall t$
 - 4: Update $\boldsymbol{\eta}^{(t)}$ in the direction of $-\partial O_{\boldsymbol{\eta}} / \partial \boldsymbol{\eta}^{(t)}, \forall t$
 - 5: **end while**
-

Following Kandemir et. al. [25], we use two different regularization functions. The first, $\Omega_1(\cdot)$, penalizes the negative total correlation, as measured by the dot product between the two kernel weight vectors $\langle \boldsymbol{\eta}^{(t_1)}, \boldsymbol{\eta}^{(t_2)} \rangle$:

$$\Omega_1(\{\boldsymbol{\eta}^{(t)}\}_{t=1}^T) = - \sum_{t_1=1}^T \sum_{t_2=1}^T \langle \boldsymbol{\eta}^{(t_1)}, \boldsymbol{\eta}^{(t_2)} \rangle$$

The second regularization function, $\Omega_2(\cdot)$, penalizes the distance of kernel weights in Euclidean space:

$$\Omega_2(\{\boldsymbol{\eta}^{(t)}\}_{t=1}^T) = \sum_{t_1=1}^T \sum_{t_2=1}^T \|\boldsymbol{\eta}^{(t_1)} - \boldsymbol{\eta}^{(t_2)}\|_2$$

3.3 Hierarchical Bayesian Logistic Regression (HBLR)

The methods we have presented so far rely on clustering participants *a priori* based on their personality and demographics, in order to build a robust model that can generalize to new people. However, it would be preferable if we could train a model to automatically cluster participants, not based on characteristics we *assume* to be related to mood prediction, but instead directly using the unique relationship each person has between their physiology, behavior, the weather, and their resulting mood. As mentioned previously, individuals may be affected very differently by the same stimuli; e.g., one person may become more calm when the weather is rainy, while another may become annoyed. The ability to group individuals based on these differing reactions could thus be highly valuable.

Therefore, we now consider a non-parametric hierarchical Bayesian model which can implicitly learn to cluster participants that are most similar in terms of their relationship between the input features and their resulting mood. Further, the model learns a soft clustering, so that a participant does not need to be assigned to a discrete, categorical cluster, but rather can belong to many clusters in varying degrees.

In hierarchical Bayesian MTL approaches, the model for each task draws its parameters from a common prior distribution. As the model is trained, the common prior is updated, allowing information to be shared across tasks. The model we adopt, which was originally proposed by Xue et. al. [32], draws logistic regression (LR) weights for each

task from a shared Dirichlet Process (DP) prior; we call this model Hierarchical Bayesian Logistic Regression (HBLR).

In contrast with our prior approaches (MTL-NN and MTMKL), the HBLR model allows us to directly define each task as predicting the wellbeing of a single person, since the model is able to implicitly learn its own clustering over people. While the implicit clustering provides valuable insights into groups of people that have a different relationship between their physiology, behavior, and wellbeing, it also means that HBLR cannot make predictions about a new person's mood without first receiving at least one labeled training data point from that person. Still, HBLR can quickly be adapted to make predictions about a new person [32], and the predictions will improve with more data.

The implicit clustering mechanism is accomplished through the choice of the Dirichlet Process prior. The DP prior induces a partitioning of the LR weights into K clusters, such that similar tasks will end up sharing the same weights. Specifically, for each task t , the model parameters $w^{(t)}$ are drawn from a common prior G which is sampled from a DP:

$$w^{(t)}|G \sim G, \quad \alpha \sim Ga(\tau_1, \tau_2) \\ G \sim DP(\alpha, G_0), \quad G_0 \sim N_d(\mu, \Sigma)$$

where Ga is a Gamma distribution and N_d is a d -dimensional multivariate normal distribution. The distribution G_0 is the base distribution and represents our prior belief about the distribution from which the weights are drawn. Following [32], we set $\mu = \mathbf{0}$ and $\Sigma = \sigma\mathbf{I}$, which reflects the prior belief that the weights should be uncorrelated and centered around zero (equally likely to be positive or negative). Here, σ is a hyperparameter. The scaling or innovation parameter of the DP $\alpha > 0$ affects the likelihood that a new cluster will be generated; as α decreases the weights generated by the DP will become more concentrated around only a few distinct clusters. In this case, α is distributed according to a diffuse prior represented by a Gamma distribution with hyperparameters τ_1 and τ_2 .

The goal of the HBLR model is to learn a posterior distribution over the variables defined above given the observed data. When each task is defined as learning the decision boundary for a single person, learning the posterior allows the model to:

- (a) learn a non-parametric clustering of similar people
- (b) perform MTL by jointly learning logistic regression classifiers for each cluster.

Here, we define people as similar when the classification boundaries of their wellbeing prediction tasks are close; that is, when their respective weight vectors are similar. This implies that similar people have a similar relationship between their input features and their resulting wellbeing.

Learning the complete posterior distribution is intractable, so mean-field variational Bayesian inference (VI) is used to approximate the true posterior; the VI equations are derived by Xue and colleagues [32]. The variational approximation of the posterior contains three sets of parameters that the model must learn. The first is a matrix $\Phi \in \mathbb{R}^{T \times K}$, where T is the number of tasks (or participants), and K is the number of clusters. The Φ is essentially the learned

soft clustering of users (see (a) above); each row $\phi^{(t)} \in R^K$ represents the degree to which person t belongs to each of the K clusters. Although the non-parametric nature of the model could theoretically allow for an infinite number of clusters, there is no loss in generality if K is limited to the number of tasks in practice. We make an additional computational enhancement to the algorithm by removing clusters for which all entries of ϕ_k are less than machine epsilon, which allows for faster convergence.

The second set of parameters are (θ_k, Γ_k) for $k = 1, \dots, K$, which parameterize a unique distribution over the LR weights for each of the K clusters (see (b) above). That is, each cluster k draws its weights from a multivariate normal distribution as follows:

$$w_k \sim N_d(\theta_k, \Gamma_k), k = 1, \dots, K$$

Note that in expectation (θ_k, Γ_k) center around the μ and Σ parameters of the base distribution.

To learn all the parameters, we use a coordinate ascent algorithm developed by Xue et. al. [32]. The parameters $(\Phi, \{\theta_k\}_{k=1}^K, \{\Gamma_k\}_{k=1}^K)$ are initialized to their respective uniform priors; that is, each task having equal contribution to each cluster to initialize Φ and setting θ_k and Γ_k to μ and Σ for each k . Each parameter is then iteratively re-estimated until convergence.

To predict a new test sample $x_*^{(t)}$, we would ideally like to use the following equation, where we integrate over the learned distribution on the classifier's weights:

$$p(y_*^{(t)} = 1|x_*^{(t)}, \Phi, \{\theta_k\}_{k=1}^K, \{\Gamma_k\}_{k=1}^K) \\ = \sum_{k=1}^K \phi_k^{(t)} \int \sigma(w_k^{*T} x_*^{(t)}) N_d(\theta_k, \Gamma_k) dw_k^*$$

where σ is the sigmoid function of a typical LR classifier. However, computing this integral is intractable. Therefore, the prediction function uses an approximate form of the integral derived in [33]:

$$p(y_*^{(t)} = 1|x_*^{(t)}, \Phi, \{\theta_k\}_{k=1}^K, \{\Gamma_k\}_{k=1}^K) \\ \approx \sum_{k=1}^K \phi_k^{(t)} \sigma \left(\frac{\theta_k^T x_*^{(t)}}{\sqrt{1 + \frac{\pi}{8} x_*^{(t)T} \Gamma_k x_*^{(t)}}} \right)$$

4 MOOD PREDICTION DATASET

The data for this research were collected as part of the "SNAPSHOT Study" at MIT and Brigham and Women's Hospital, an ambulatory study of Sleep, Networks, Affect, Performance, Stress, and Health using Objective Techniques [34]. Participants were college students who were monitored for 30 days each. The study gathers rich data, including daily smartphone, physiological, behavioral and mood data.

4.1 Classification labels and dataset

The goal of this work is to use the data from the SNAPSHOT study for predicting students' wellbeing (mood, stress, and health). Each morning and evening, participants self-reported their mood (sad/happy), stress (stressed

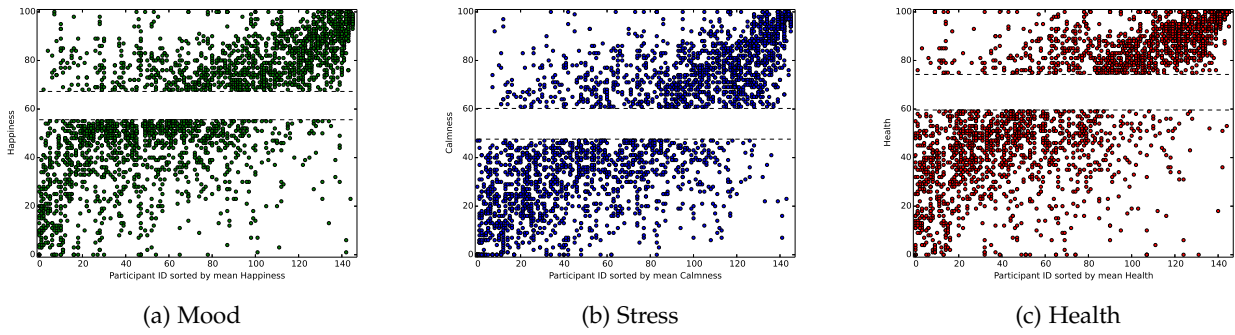


Fig. 2: Distribution of self-report labels after discarding the middle 20%. Participants are listed on the x-axis, in order of their average self-report value for that label (each participant is one column). Almost all participants have data from both label classes.

out/calm), and health (sick/healthy) on a visual analog scale from 0-100; these scores are split based on the median value to create binary classification labels. Previous work has relied on discarding the most neutral scores before creating binary labels in order to disambiguate the classification problem; i.e., the middle 40% of scores were discarded due to their questionable nature as either a ‘happy’ or ‘sad’ state [9], [10]. We instead make the problem decidedly harder by discarding only the middle 20% of scores. We also discard participants with less than 10 days worth of data, since they could not provide enough data to train viable models. The resulting dataset comprises 104 users and 1842 days.

Figure 2 shows the raw values reported for mood, stress, and health for each participant after the middle 20% of scores have been removed. Points appearing above the removed values are assigned a positive classification label, while points below are assigned a negative label. As is apparent from the figures, although some participants predominantly report one label class almost all participants’ reports span the two classes. This implies that the need for personalization is not simply due to the fact that some participants are consistently sad while some are consistently happy, for example. Personalization is required because people react differently to very similar stimuli, and a single, impersonal classifier cannot capture these differences.

4.2 Features

To predict the labels, 343 features are extracted from the smartphone logs, location data, physiological sensor recordings, and behavioral surveys obtained about participants each day. Due to the rich, multi-scale nature of the data collected, careful feature extraction is critically important, and has been explored in detail in previous work [9], [10]. Here we provide a brief overview of the feature types.

Physiology: 24-hour-a-day skin conductance (SC), skin temperature, and 3-axis acceleration were collected at 8 Hz using wrist-worn Affectiva Q sensors. SC is controlled by the sympathetic nervous system (SNS); when the body experiences a “fight or flight” response, a peak in the SC signal termed an SCR will occur. Using the SC signal, we automatically remove noise using a pre-trained algorithm [35], detect SCRs, and compute features related to their amplitude, shape, and rate, which are shown in Figure 3.

The skin temperature and accelerometer data are also used to compute features; from the latter, we extract measures of activity, step count, and stillness. Since physical activity reduces stress and improves mood [36], and skin temperature is related to the body’s circadian rhythm [37], we expect these features to be highly relevant. We also weight the SCR features by stillness and temperature, since we are interested in SCRs due to emotion and stress rather than exertion or heat. In total we compute 172 physiology features over different periods of the day.

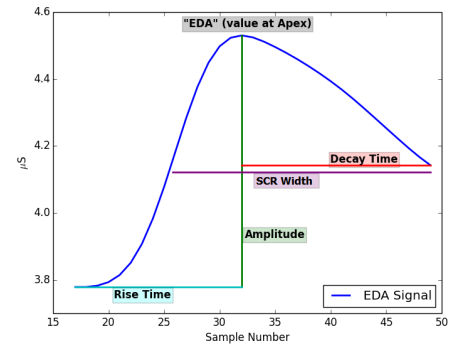


Fig. 3: Features extracted for each detected, non-artifact SCR

Phone (call, SMS, screen): an app on participants’ phones logs their calls, text messages (SMS), and whenever the phone’s screen is turned on or off. Features are computed based on the timing and duration of these events and the number of unique contacts with whom each person interacts. This gives a total of 20 call, 30 SMS, and 25 screen features. An example of SMS data is shown in Figure 4, in which the texting pattern on a sad day appears noticeably different than on another day.

For both the physiology and phone, each feature set is computed over four time intervals: 12-3AM, 3-10AM, 10AM-5PM, 5-11:59PM. These intervals were determined by examining density plots of the times students were most likely to be asleep (3-10AM), or in class (10AM-5PM), as shown in Figure 5.

Behavioral surveys and extrinsic variables: 38 features are computed on students’ self-reported extra-curricular and academic activities, exercising, sleeping, napping, social interactions, and alcohol and drug consumption. We also

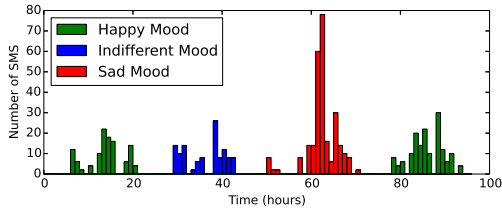


Fig. 4: SMS frequency over four days

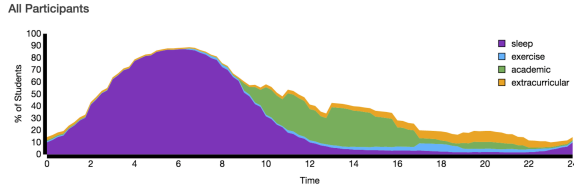


Fig. 5: Percent of participants sleeping, studying, in extra-curricular activities, and exercising throughout the day.

include 3 extrinsic variables that would be available to any smartphone app: the participant ID, the day of the week, and whether it is a school night.

Weather: Previous studies have reported on how the weather effects mood, particularly in relation to Seasonal-Affective Disorder [37], [38]. Additionally, it is well known that there are particular seasons of the year (i.e. winter) that have higher rates of poor health. Therefore, we extracted 40 features about the weather from DarkSky’s Forecast.io API [39]. These features include information about sunlight, temperature, wind, Barometric pressure, and the difference between today’s weather and the rolling average.

Location: the smartphone app logs participants’ GPS coordinates throughout the day. After cleaning, interpolating, and downsampling this signal, we compute a total of 15 features including the total distance traveled, the radius of the minimal circle enclosing the location samples, time spent on the university campus, and time spent outdoors based on wifi usage. The location coordinates are also used to learn a Gaussian Mixture Model (GMM), giving a probability distribution over each participant’s typical locations (see Figure 6). We then compute features such as the log likelihood of the location pattern for each day. In essence, this measures the routineness of the participant’s day, which we have found to be negatively associated with happiness and calmness [9]. Note that since the GMM is learned from the location history of a participant, models trained on these features cannot be run without collecting a few days of location data for each participant; still, the participant does not need to self-report labels to benefit from the model.

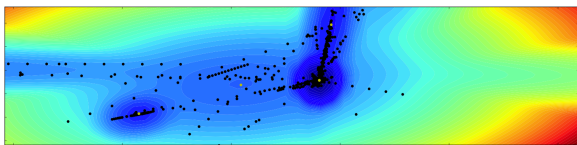


Fig. 6: GMM fitted to location data from one participant. Black points are locations visited; the contours mark the probability distribution induced by the model.

Feature selection: Since the dataset is small, we apply feature selection to reduce the chance of overfitting. While there are many ways to do this, in this work features are selected based on assessing ANOVA F-scores between each feature and the classification label using the training data and removing highly correlated features, with the constraint that at least one feature from each of the above data sources is retained. This process gave rise to a total of 21 features, which are listed in Table 1.

TABLE 1: Selected 21 features and modalities

Modality	Features
Classifier	Day of the week
Physiology 3am-10am	% mins with ≥ 5 SCRs (w/o artifacts) Temperature weighted SCR AUC
Location	Time on campus Log likelihood of day given previous days
Call	Total missed calls
SMS	Total incoming (midnight-3am) Number of unique contacts outgoing Number unique incoming (5-11:59pm) Number unique outgoing (5-11:59pm)
Screen	Total duration (Midnight-3am) Total number on/off events (5-11:59pm)
Survey Activities	Exercise duration Study duration
Survey Interaction	Positive social interaction Presleep in-person interaction (T/F)
Survey Sleep	Number of naps All-nighter (T/F)
Weather	Cloud cover rolling std. dev. Max precipitation intensity Pressure rolling std. dev.

4.3 Pre-study survey data

At the beginning of the SNAPSHOT study participants completed personality and mental health inventories. These measures included Myers-Briggs and Big Five Factor personality scores, state and trait anxiety scores, the Short-Form 12 Mental health Composite Score (MCS), Physical health Composite Score (PCS), Pittsburgh Sleep Quality Index (PSQI), the Perceived Stress Scale (PSS) and the participant’s GPA and BMI (see [34] for details on these measures). While this data is not incorporated directly into the MTL models (except through the K -means clusters described in Section 3), we hypothesize that it may be relevant to the soft clustering learned by HBLR.

5 EXPERIMENTS

To assess whether personalization via MTL provides significant performance benefits, we compare it to two other approaches. First, we compare each algorithm to its single task learning (STL) equivalent. HBLR is compared to conventional LR, MTMKL to LSSVM, and MTL-NN to a generic NN. Second, to determine whether personalization via MTL has a performance advantage over simply using MTL itself, we also explore multitasking over the related wellbeing measures; in other words, in this condition we treat predicting mood, stress, and health as related tasks. Note that this moods-as-tasks approach to MTL is similar to that taken in prior work (e.g. [10], [25], [27], [28]).

To create the datasets used for training the models and avoid testing data contamination, a random 80/20% split

was used to partition the SNAPSHOT data into a train and test set. We then apply 5-fold cross validation to the training set to test a number of different parameter settings for each of the algorithms described in Section 3. Finally, each model is re-trained using the optimal parameter settings, and tested once on the held-out testing data; the test performance is reported in the following section.

Due to space constraints and the number of models investigated, we do not report the optimal hyperparameter settings for each model but will provide them upon request. Instead, we will simply specify that for training the NNs, we consistently used learning rate decay and the Adam optimizer [40], and tuned the following settings: the number and size of hidden layers, batch size, learning rate, whether or not to apply dropout, and the L2 β weight. Based on previous work that has successfully trained MTL NNs with few samples [22], we choose a simple, fully-connected design with 2-4 hidden layers. For HBLR, we tuned the τ_1 , τ_2 , and σ parameters, while for MTMKL we tuned C , β , the type of kernel (linear vs. radial basis function (RBF)), the type of regularizer function ($\Omega_1(\cdot)$ vs $\Omega_2(\cdot)$), and ν . For MTMKL we also define the following modalities: classifier, location, survey interaction, survey activities, survey sleep, weather, call, physiology from 3am to 10am, screen, and SMS. More detail on these modalities is provided in Table 1.

All of the code for the project, which is written in Python and TensorFlow [41], has been released open-source and is available at <https://github.com/mitmedialab/personalizedmultitasklearning>

5.1 Analysis of HBLR clusters

Because the clusters learned by the HBLR model may be fundamentally different than those that can be obtained using other methods, we are interested in defining a way to analyze which type of participants are represented within each cluster. For example, does a certain cluster tend to contain participants that have a significantly higher trait anxiety score (as measured by the pre-study survey)?

The analysis is complicated by the fact there is no discrete assignment of participants to clusters; rather, a participant may have some degree of membership in many or all of the clusters, as defined by $\phi^{(t)}$. To solve this issue, we first define a matrix $\mathbf{P} \in \mathbb{R}^{T \times M}$, where T is the number of participants and M is the number of pre-study measures (such as Big Five personality, PSS, etc.). Thus, $P_{t,m}$ represents person t 's score on measure m . Using \mathbf{P} , we can then compute a score representing the average value of each pre-study measure for each cluster, as follows:

$$Q_{k,m} = \frac{\sum_t P_{t,m} \phi_k^{(t)}}{\sum_t \phi_k^{(t)}}$$

where $\mathbf{Q} \in \mathbb{R}^{K \times M}$ and K is the number of clusters learned by the HBLR model. $Q_{k,m}$ can be considered a weighted average of a cluster's pre-study trait, where the weights are the degree of membership of each participant in that cluster.

To test whether a cluster's $Q_{k,m}$ value is significantly different than the group average, we use a one-samples t-test to compare $Q_{k,m}$ to the values for measure m reported by participants on the pre-study survey. We apply a Bonferroni correction based on the number of comparisons made

across the different clusters within each outcome label (i.e. mood, stress, health).

6 RESULTS AND DISCUSSION

The accuracy and Area Under the ROC Curve (AUC) of each of the mood prediction models is shown in Table 2, along with the majority class baseline (the performance that can be expected from simply predicting the most frequent label in the training data). For most models, we found that using feature selection improved performance. Since NNs often benefit from large input vectors, we tested the performance of the MTL-NN on the full set of 343 features as well, and include these results in Table 2.

As shown in the table, the accuracy obtained using traditional STL ML classifiers is poor, reaching a maximum of only 60-66%; this is similar to prior work that trained STL classifiers to detect mood on a simplified version of this dataset [9]. The performance obtained with the three MTL models when multitasking over the related outcome labels, i.e. mood, stress, and health is shown as *MTL - moods*. Evidently, multitasking in this way does not significantly enhance performance. This could be because the outcome labels are not sufficiently related to each other to benefit from sharing parameters, or that the regularization imposed by MTL limits the models' capacity more than it benefits the generalization performance. Therefore, it is clear that at least for this data, MTL alone is not sufficient to improve mood prediction classifiers.

Rather, it is using MTL to account for individual differences that is important. As is clear from both Table 2 and Figure 7, using MTL to personalize ML models by multitasking over clusters of similar people provides dramatic improvements to mood prediction performance. The improvement in accuracy over the non-personalized models ranges from 11-21%. McNemar tests of the predictions with a Bonferroni correction applied within each label type revealed that the personalized models significantly outperformed ($p < .05$) both the STL and *MTL - moods* approaches, over all model and label types. These scores represent state-of-the-art performance on this dataset, surpassing prior published work by 5-13% prediction accuracy [9], [10].

	Classifier	Mood	Stress	Health
Baseline	Majority class	50.4%, .500	50.7%, .500	54.4%, .500
STL	LSSVM	60.2%, .603	58.1%, .581	62.3%, .614
	LR	56.9%, .569	59.4%, .594	55.4%, .544
	NN	60.5%, .606	60.1%, .600	65.9%, .648
	NN (all feats)	65.8%, .658	67.9%, .678	59.0%, .591
MTL - moods	MTMKL	59.4%, .594	58.8%, .587	62.0%, .610
	HBLR	58.3%, .583	57.8%, .578	55.1%, .551
	MTL-NN	60.2%, .602	60.1%, .600	65.3%, .643
	MTL-NN (all feats)	67.0%, .670	68.2%, .682	63.0%, .623
MTL - people	MTMKL	78.7%, .787	77.6%, .776	78.7%, .786
	HBLR	72.0%, .720	73.4%, .734	76.1%, .760
	MTL-NN	77.6%, .776	78.6%, .785	79.7%, .792
	MTL-NN (all feats)	78.4%, .784	81.5%, .815	82.2%, .818

TABLE 2: Prediction performance (Accuracy and AUC) of the STL, MTL-moods, and MTL-user methods. Bolded entries represent significant improvements over the STL model, indicating that multitasking for personalization is by far the most effective approach.

Given the impressive performance of the personalized MTL models, in the following sections we focus on analyzing the weights and clusters learned by the personalized

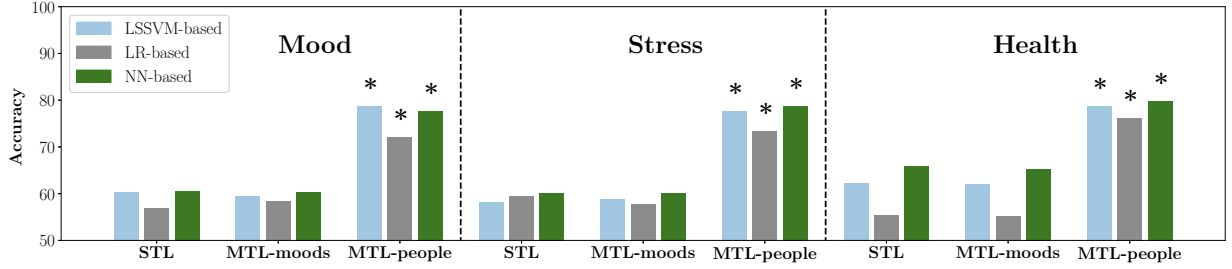


Fig. 7: Accuracy for each type of model in the STL, MTL-moods, and MTL-people approaches. Note that the accuracy significantly ($* = p < 0.05$) improves when using multi-tasking over people for each label and for each machine learning method tested.

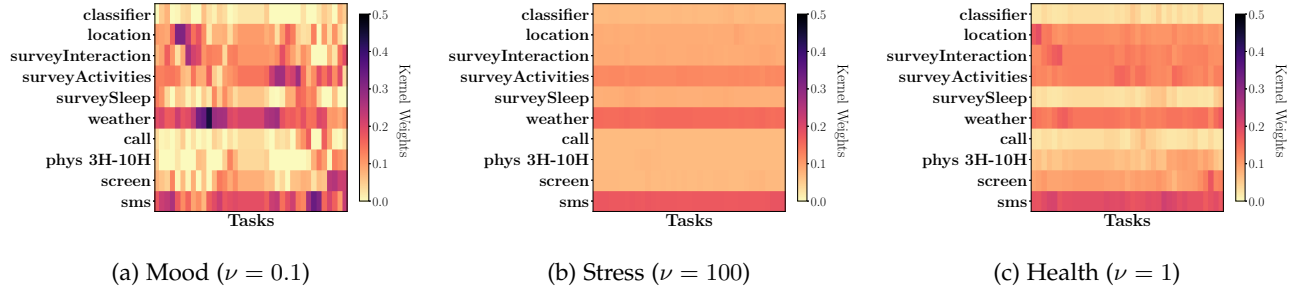


Fig. 8: MTMKL kernel modality weights, reflecting which feature type is most important to the classifier for each task. The ν parameter controls how heavily the task weights are regularized to be similar, and was set by the hyperparameter search.

MTMKL and HBLR models, both of which can help provide important insights into how the wellbeing of different groups of people is affected by their physiology, their behavior, their use of technology, and the weather.

6.1 MTMKL

The MTMKL model learns a weighting over the 10 modalities for each task. As described in Section 3.2, the ν parameter controls how strongly the tasks' weights are constrained to be similar. Figure 8 shows the weights learned by the personalized MTMKL model for each outcome label and each cluster of people. The figure demonstrates that for predicting stress, the hyperparameter search selected a large value of ν , which constrained the kernel weights to be highly similar across all clusters. However, for mood the value of ν was much smaller, resulting in more diverse kernel weights for the different tasks. This could suggest that there is more individual variability in how well these features can predict mood, while the relationship to stress is more consistent. It appears that for mood prediction, differences in features like location and screen are important for some types of people but not others.

The modality weights learned by MTMKL can potentially provide interesting insights for designing a mood prediction system. For example, we see that for this set of features, overall the differences in weather, SMS, and survey (representing exercise and studying duration) tend to be more informative. This may suggest that not only is it important to include weather in a mood prediction system, but that developing ways to automatically and unobtrusively detect when a participant is exercising or studying

could be a valuable time investment. Further, we see that the call features tend to be less informative compared to the other features, so perhaps it is not necessary to monitor participants' call patterns to predict wellbeing. Removing this form of data collection could potentially enhance privacy for participants in the SNAPSHOT study.

6.2 HBLR

The HBLR model learns a non-parametric soft clustering of participants based on the relationship between their input features and resulting mood. Figure 9 shows the clustering learned for predicting each of the three outcome labels, where the intensity of the color represents the degree to which each participant belongs to each cluster. The number of clusters which had at least one participant with a degree of membership exceeding machine epsilon were 4, 3, and 17 for the mood, stress, and health prediction models, respectively. However, this does not imply that there are only three types of people which have a different relationship between the features and stress. Because of the soft clustering, a given person can belong to many clusters and thus combine the decision boundaries learned for each, as explained below.

As discussed previously, each cluster in the HBLR model learns a multivariate normal distribution over the weight vector w_k^* . In Figure 10 we show examples of the different marginal distributions learned over a single feature (total number of screen on events (5pm-midnight)) for the four mood clusters. We note that for these two features, cluster 0 and cluster 1 have very different distributions on the LR weights. For example, in Figure 10 we see that cluster 0 places a negative weight on the feature whereas cluster 1

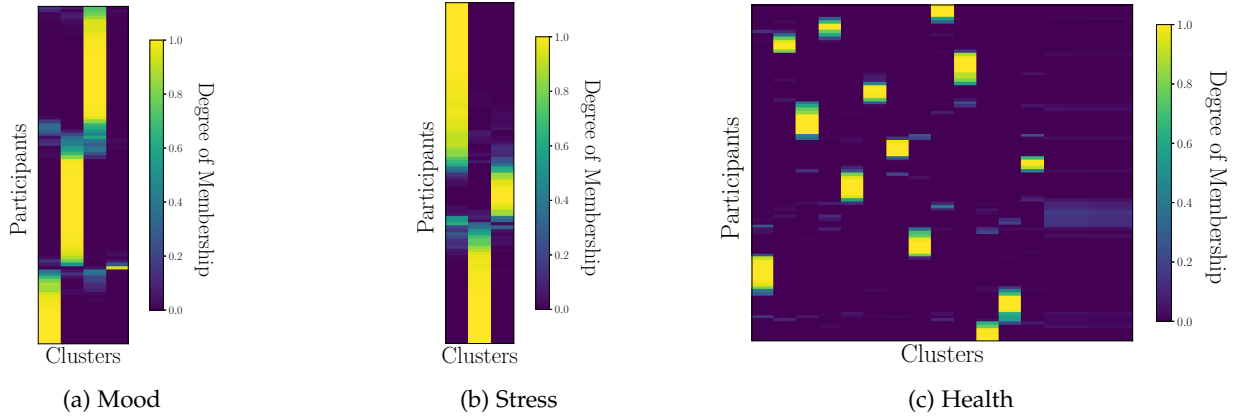


Fig. 9: Resulting soft clustering (Φ) when predicting the different labels (mood, stress, and health). Each row shows one of the 104 participant's degree of membership in each cluster. We note that there were 4,3, and 17 clusters needed in predicting happiness, stress, and health, respectively.

places a positive weight on the same feature. Thus, when participants who belong almost exclusively to cluster 0 use their phone excessively in the evening, the model will be more likely to predict a sad day tomorrow. In contrast, the model is more likely to predict a happy day tomorrow for participants belonging almost exclusively to cluster 1 based on the same behavior.

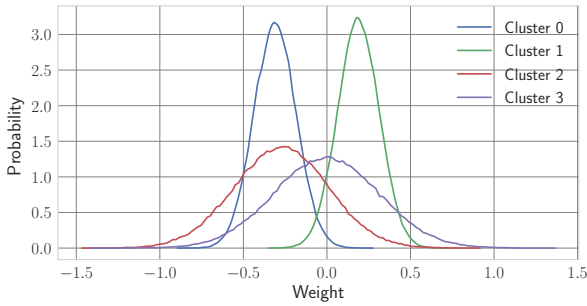


Fig. 10: Distribution of HBLR weights on the *total number of screen on events (5pm-midnight)* feature for each cluster when predicting tomorrow's mood

However, because participants do not belong exclusively to one cluster or another, the marginal distribution over a weight parameter for a given participant can be more complex than a multivariate normal. For example, Figure 11 shows an example of the weight distributions for 3 different participants. For Participant 5, the model has constructed a bimodal distribution over the weight by combining the distributions of multiple clusters. Thus, the model is able to customize the decision boundary for each person while still clustering the participants into similar archetypes.

As described in Section 5.1, we would like to determine if the clusters learned by the HBLR model differ significantly in terms of the typical personality or mental health scores of the participants. Following the procedure outlined in that section, we computed the average scores for each cluster on each of the pre-study trait measures (i.e. the matrix \mathbf{Q}), then conducted a limited number of significance tests with a Bonferroni correction to determine if there were significant

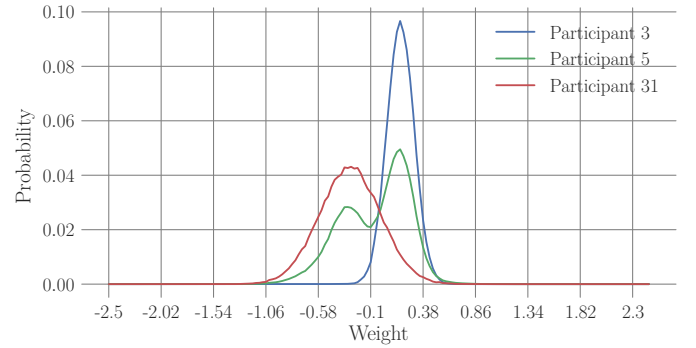


Fig. 11: Example of different weight distributions induced by the soft clustering for 3 different participants in the mood prediction. Participant 3 is almost exclusively in cluster 1, participant 5 is has membership in clusters 0, 1, and 2, and participant 31 is almost exclusively in cluster 2.

differences among the clusters for some of the traits. Since the HBLR clustering is based on latent factors underlying the data that are unknown before training, it is not possible to determine *a priori* what traits may be particularly relevant to a given cluster. Below, we discuss the results of these computations for some notable traits of the mood and stress clusters. We do not show the same analysis for health, since the 17 different clusters in the health model render it impractical to present the results.

Table 3 shows the relevant trait values for the mood clusters, including the average value for those traits computed over all participants in the study. According to these findings, the clusters learned by the HBLR model in predicting mood can be characterized as a) Judging and Sensing personality types; b) people with better than average sleep quality (PSQI); c) Agreeable people, and d) happy Extraverts with low state and trait anxiety. This could suggest that these traits are highly relevant for predicting how a person's mood will change given the input features. For example, since poor sleep quality has been shown to have a negative effect on mood [42], perhaps the normally high sleep quality of participants in cluster 1 makes their mood more sensitive

to sleep disturbances.

Cluster	Pre-study measure	All participants	Cluster $Q_{k,m}$	t	p
0	Percent happy days	$M = 49, SD = 37$	56	-1.86	> .10
0	Judging	$M = 61, SD = 21$	73	-7.69	< .001
0	Sensing	$M = 47, SD = 20$	57	-7.22	< .001
1	Percent happy days	$M = 49, SD = 37$	55	-1.81	> .10
1	PSQI	$M = 4.7, SD = 2.3$	4.1	3.48	< .01
2	Percent happy days	$M = 49, SD = 37$	41	2.29	> .10
2	Agreeableness	$M = 50, SD = 28$	43	3.63	< .01
3	Percent happy days	$M = 49, SD = 37$	78	-8.00	< .001
3	Extraversion	$M = 49, SD = 30$	76	-13.1	< .001
3	State anxiety	$M = 38, SD = 10$	30	10.9	< .001
3	Trait anxiety	$M = 43, SD = 10$	36	9.85	< .001

TABLE 3: Computed pre-study measures for the HBLR mood prediction clusters. Bolded entries represent significant differences from the sample average.

It is particularly interesting to relate these results to the average value for the weights learned for these clusters, as shown in Figure 12. For example, it appears that the “Agreeable” cluster (cluster 2) places high weight on four social interaction features; this is consistent with research indicating that those with an Agreeable personality type value getting along with others [29]. In contrast with this cluster, the “High sleep quality” cluster (cluster 1) places negative weight on features related to SMS use in the evening. Finally, we observe that the “Judging and Sensing” cluster (cluster 0) has a positive association with exercise, but a negative association with time spent on campus.

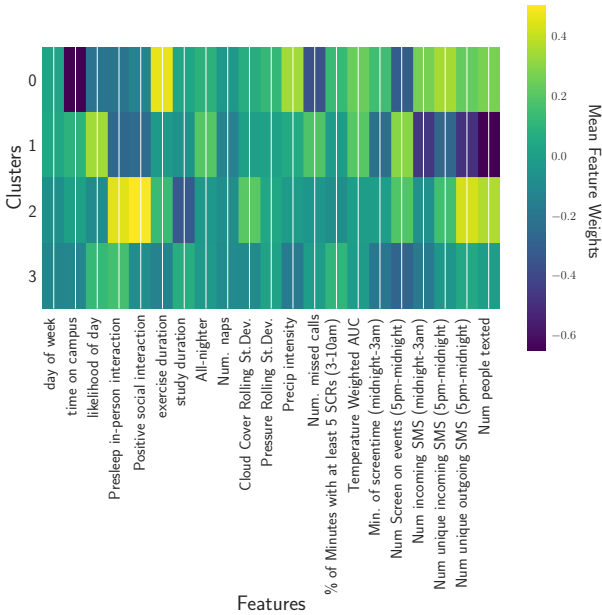


Fig. 12: Mean feature weights for mood clusters in HBLR model. We note that the positive label is “Happy” so features with positive (negative) mean weights contribute to being more happy (sad) tomorrow.

Note that we also examined whether the HBLR model would simply cluster participants with a tendency to be particularly happy or particularly sad together, in order to more easily make accurate predictions. As shown in Table 3, three of the clusters do not differ significantly from the group average in terms of average percent of happy days, although cluster 3 (Extroverts with low state and trait anxiety) does correspond to particularly happy participants.

The results of the same analysis of HBLR cluster pre-study measures for the stress model are shown in Table 4. In this case, none of the clusters differed significantly from the group average in terms of the percentage of calm days. While we did not detect any significant differences from the group average for cluster 0, cluster 1 represents an intuitively salient group: conscientious people with a high GPA. It makes sense that this clustering would be relevant to predicting stress, since conscientious students who are concerned about their grades are likely to have strong stress reactions in an academic environment. As shown in Figure 13 this cluster places a positive weight on the “likelihood of day” feature, which is a measure of how routine the participants location patterns were that day, and will be higher if the participant travels mainly to typical work and home locations. Stress cluster 2 represents students who are extraverted, with slightly increased BMI and lowered physical health. In examining Figure 13, we can see that cluster 2 has highly positive mean feature weights on the SMS features, which is consistent with the trait of Extraversion. On the contrary, cluster 1 has highly negative weights on the social SMS features, meaning more SMS use for these participants would increase the likelihood of predicting a stressful day tomorrow. One of several possible explanations is that perhaps these conscientious, high GPA students become stressed by having to balance their academic goals and social life.

Cluster	Pre-study measure	All participants	Cluster $Q_{k,m}$	t	p
0	Percent calm days	$M = 48, SD = 38$	46	.492	> .60
1	Percent calm days	$M = 48, SD = 38$	55	-1.88	> .10
1	GPA	$M = 4.4, SD = .61$	4.6	-3.95	< .001
1	Conscientiousness	$M = 51, SD = 28$	58	-3.43	< .01
2	Percent calm days	$M = 48, SD = 38$	39	2.32	> .10
2	Extraversion	$M = 49, SD = 30$	58	-4.50	< .001
2	BMI	$M = 24, SD = 4.4$	25	-4.09	< .001
2	PCS	$M = 58, SD = 4.2$	57	3.77	< .01

TABLE 4: Computed pre-study measures for the HBLR stress prediction clusters. Bolded entries represent significant differences from the sample average.

7 CONCLUSIONS AND FUTURE WORK

This work has demonstrated that accounting for individual differences through MTL can substantially improve mood and wellbeing prediction performance. This performance enhancement is not simply due to the application of MTL, but rather through the ability of MTL to allow each person to have a model customized for them, but still benefit from the data of other people through hidden layers of a deep neural network, kernel weights, or a shared prior. The three methods we have explored offer different strengths, including the ability to learn how the importance of feature types differs across tasks, and the ability to learn implicit groupings of users.

7.1 Limitations and Future Work

A major limitation of this research relates to the relatively small sample size. With data from more individuals and with longer monitoring per person, it may be possible to build time series models for forecasting, which could be even more accurate and powerful if personalized. In the ideal case, we would like to be able to predict a person’s

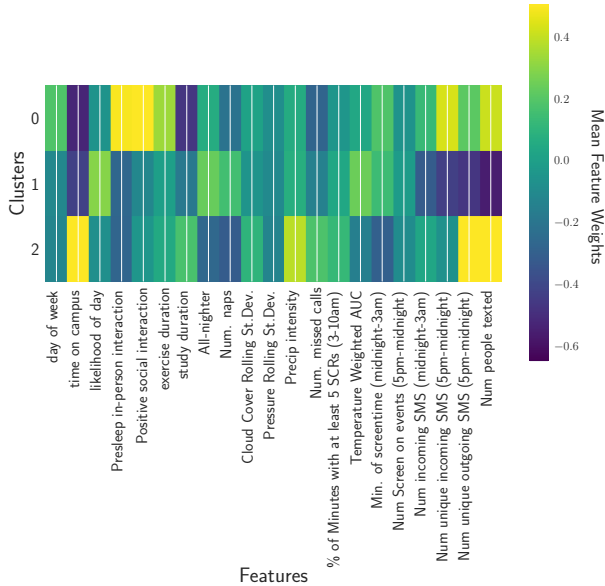


Fig. 13: Mean feature weights for stress clusters in HBLR model. We note that the positive label is “Calm” so features with positive (negative) mean weights contribute to being more calm (stressed) tomorrow.

wellbeing far into the future, rather than just one day in advance. Given the scope of the work undertaken here, there are many other aspects that could be enhanced. For example, the features were selected using a generic method that did not take the classifier type into consideration. We also do not use several features suggested by other work as important for mental health and wellbeing [34]. Future work will more deeply examine these cases and their interplay with a personal model. Further, we can also incorporate exciting new techniques for improving MTL in future, for example by combining the hierarchical Bayesian approach with deep neural networks (DNNs) [43], or training DNNs that are able to quickly adapt parameters in order to learn new tasks [44].

It is worth emphasizing that the presented algorithms could provide the ability to make predictions about a novel person who has not provided any self-report labels. If this person is willing to complete a personality inventory, predictions can be made immediately using the MTL-NN and MTMKL models, which are based on K-means clusters computed from personality and gender data. The HBLR model can be extended to make mood predictions for a novel user who has not provided classification labels, by applying MCMC to her data [32]. In future work, we will assess the classification accuracy of these models on novel participants. A strongly motivating application goal is to be able to detect individuals with low wellbeing or mental health in order to guide prevention or early intervention efforts. If new users were only required to install an app that did not depend on them inputting mood data, interventions would be able to reach a larger population.

Another aim of this research is to generate hypotheses about problematic behaviors that are indicative of low mood and mental health. Through examination of the model weights and clusters, we hope to gain insight into the be-

haviors that are significant wellbeing predictors for people with different personalities and lifestyles. Once hypotheses related to these behaviors have been refined, we can test them via causal inference techniques such as counterfactual reasoning. These inferences would be useful for anyone wishing to know what types of behaviors best promote a happy, calm, and healthy state.

Finally, we hope that by providing the code for these techniques, other authors will be encouraged to use them to personalize models for a wide variety of problems in which interindividual variability is important. When used in conjunction with the analysis techniques outlined here, these models may not only lead to the discovery of interesting insights across many problems, but may help to significantly enhance performance in predicting difficult, ambiguous, and personal outcomes.

ACKNOWLEDGMENTS

We would like to thank Dr. Charles Czeisler, Dr. Elizabeth Klernman, and other SNAPSHOT project members for their help in designing and running the SNAPSHOT study. This work was supported by the MIT Media Lab Consortium, NIH Grant R01GM105018, Samsung Electronics, NEC Corporation, and Canada’s NSERC program.

REFERENCES

- [1] H. Cheng and A. Furnham, “Personality, self-esteem, and demographic predictions of happiness and depression,” *Personality and individual differences*, vol. 34, no. 6, pp. 921–942, 2003.
- [2] R. Veenhoven, “Healthy happiness: Effects of happiness on physical health and the consequences for preventive health care,” *Journal of happiness studies*, vol. 9, no. 3, pp. 449–469, 2008.
- [3] S. Cohen *et al.*, “Psychological stress and susceptibility to the common cold,” *New England journal of medicine*, vol. 325, no. 9, pp. 606–612, 1991.
- [4] A. Keller *et al.*, “Does the perception that stress affects health matter? the association with health and mortality,” *Health Psychology*, vol. 31, no. 5, p. 677, 2012.
- [5] S. Aichele, P. Rabbitt, and P. Ghisletta, “Think fast, feel fine, live long: A 29-year study of cognition, health, and survival in middle-aged and older adults,” *Psychological science*, vol. 27, no. 4, pp. 518–529, 2016.
- [6] A. Bogomolov *et al.*, “Daily stress recognition from mobile phone data, weather conditions and individual traits,” in *Int. Conf. on Multimedia*. ACM, 2014, pp. 477–486.
- [7] R. LiKamWa *et al.*, “Moodscope: building a mood sensor from smartphone usage patterns,” in *Int. Conf. on Mobile systems, applications, and services*. ACM, 2013, pp. 389–402.
- [8] A. Grunerbl, A. Muaremi, V. Osmani, G. Bahle, S. Oehler, G. Tröster, O. Mayora, C. Haring, and P. Lukowicz, “Smartphone-based recognition of states and state changes in bipolar disorder patients,” *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 140–148, 2015.
- [9] N. Jaques *et al.*, “Predicting students’ happiness from physiology, phone, mobility, and behavioral data,” in *ACII*. IEEE, 2015.
- [10] —, “Multi-task, multi-kernel learning for estimating individual wellbeing,” in *NIPS 2015 Workshop on Multimodal ML*, vol. 898, 2015.
- [11] J. Brebner, “Personality factors in stress and anxiety,” *Cross-cultural anxiety*, vol. 4, pp. 11–19, 1990.
- [12] L. Clark, D. Watson, and S. Mineka, “Temperament, personality, and the mood and anxiety disorders,” *Journal of abnormal psychology*, vol. 103, no. 1, p. 103, 1994.
- [13] T. Klimstra *et al.*, “Come rain or come shine: individual differences in how weather affects mood,” *Emotion*, vol. 11, no. 6, p. 1495, 2011.
- [14] R. Caruana, “Multitask learning,” *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

- [15] A. Bogomolov *et al.*, "Happiness recognition from mobile phone data," in *Social Computing (SocialCom), 2013 International Conference on*. IEEE, 2013, pp. 790–795.
- [16] D. Carneiro, P. Novais, J. C. Augusto, and N. Payne, "New methods for stress assessment and monitoring at the workplace," *IEEE Transactions on Affective Computing*, 2017.
- [17] S. Koldijk, M. A. Neerincx, and W. Kraaij, "Detecting work stress in offices by combining unobtrusive sensors," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2017.
- [18] L. Canzian and M. Musolesi, "Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 2015, pp. 1293–1304.
- [19] Y. Suhara, Y. Xu, and A. Pentland, "Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 715–724.
- [20] M. Rosenstein *et al.*, "To transfer or not to transfer," in *NIPS 2005 Workshop on Transfer Learning*, vol. 898, pp. 1–4.
- [21] C. Zhang and Z. Zhang, "Improving multiview face detection with multi-task deep convolutional neural networks," in *Applications of Computer Vision*. IEEE, 2014, pp. 1036–1041.
- [22] F. Jin and S. Sun, "Neural network multitask learning for traffic flow forecasting," in *IEEE Int'l Joint Conf. on Neural Networks*. IEEE, 2008, pp. 1897–1901.
- [23] J. Baxter, "A bayesian/information theoretic model of learning to learn via multiple task sampling," *Machine Learning*, vol. 28, no. 1, pp. 7–39, 1997.
- [24] A. Wilson *et al.*, "Multi-task reinforcement learning: a hierarchical bayesian approach," in *ICML*. ACM, 2007, pp. 1015–1022.
- [25] M. Kandemir *et al.*, "Multi-task and multi-view learning of user state," *Neurocomputing*, vol. 139, pp. 97–106, 2014.
- [26] S. Koelstra *et al.*, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.
- [27] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on Affective Computing*, 2015.
- [28] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences," *IEEE Transactions on Affective Computing*, 2017.
- [29] O. P. John and S. Srivastava, "The big five trait taxonomy: History, measurement, and theoretical perspectives," *Handbook of personality: Theory and research*, vol. 2, no. 1999, pp. 102–138, 1999.
- [30] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.
- [31] N. Srivastava *et al.*, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [32] Y. Xue *et al.*, "Multi-task learning for classification with dirichlet process priors," *The Journal of Machine Learning Research*, vol. 8, pp. 35–63, 2007.
- [33] D. J. MacKay, "The evidence framework applied to classification networks," *Neural computation*, vol. 4, no. 5, pp. 720–736, 1992.
- [34] A. Sano, "Measuring college students sleep, stress and mental health with wearable sensors and mobile phones," Ph.D. dissertation, MIT, 2015.
- [35] S. Taylor *et al.*, "Automatic identification of artifacts in electrodermal activity data," in *EMBC*. IEEE, 2015.
- [36] J. Ratey, *Spark: The revolutionary new science of exercise and the brain*. Hachette Digital, Inc., 2008.
- [37] T. Partonen, "Dopamine and circadian rhythms in seasonal affective disorder," *Medical hypotheses*, vol. 47, no. 3, pp. 191–192, 1996.
- [38] J. Li, X. Wang, and E. Hovy, "What a nasty day: Exploring mood-weather relationship from twitter," in *Int'l Conf. on Info. and Knowledge Management*. ACM, 2014, pp. 1309–1318.
- [39] T. D. S. C. LLC. (2016) Dark sky forecast api. [Online]. Available: <https://developer.forecast.io/>
- [40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: <http://arxiv.org/abs/1412.6980>
- [41] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: <http://tensorflow.org/>
- [42] B. Bower *et al.*, "Poor reported sleep quality predicts low positive affect in daily life among healthy and mood-disordered persons," *Journal of sleep research*, vol. 19, no. 2, pp. 323–332, 2010.
- [43] R. Salakhutdinov, J. Tenenbaum, and A. Torralba, "Learning with hierarchical-deep models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1958–1971, 2013.
- [44] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *CoRR*, vol. abs/1703.03400, 2017. [Online]. Available: <http://arxiv.org/abs/1703.03400>



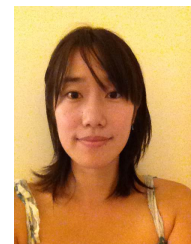
Sara Taylor is a Ph.D. student in the Affective Computing group at the MIT Media Lab, where she studies how to use physiological and behavioral data to predict mood and health. She has also worked on new ways to collect self-reported data from participants in an unobtrusive way. Sara has a Master's in Media Arts and Sciences from MIT, and a B.S. in Electrical Engineering and a B.S. in Mathematics both from Brigham Young University.



Natasha Jaques is a Ph.D. student in the Affective Computing group at the MIT Media Lab, studying machine learning and deep learning techniques in service of predicting and interpreting psychological states. Natasha has a M.Sc. in Computer Science from the University of British Columbia, and graduated from the University of Regina with an Honours B.Sc. in Computer Science and a B.A. in Psychology. She has received numerous awards, including the NIPS 2016 Best Demo, RWJF Wellbeing Fellowship, MSR Graduate Womens Scholarship, UBC CS Merit Scholarship, and the S.E. Stewart Award in Arts.



Ehimwenma Nosakhare is a Ph.D. candidate at MIT's Department of Electrical Engineering and Computer Science, and the Affective Computing group at the MIT Media Lab. Her research interests are in Bayesian modeling and causal inference. She is particularly interested in using these techniques to improve mood, mental health and wellbeing. She received her S.M. in Electrical Engineering and Computer Science from MIT, and a B.Sc. in Electrical Engineering, *summa cum laude*, from Howard University.



Akane Sano is a research scientist in the Affective Computing Group at the MIT Media Lab. Her research focuses on multi-modal ambulatory human sensing using wearable sensors and mobile phones, data analysis/modeling and application development for affective computing, health and wellbeing. She received her Ph.D. from MIT and an M.Eng. and a B.Eng. from Keio University, Japan.



Rosalind Picard, ScD, FIEEE, is founder and director of the Affective Computing Research Group at the Massachusetts Institute of Technology (MIT) Media Laboratory, co-founder of Affectiva, and co-founder and Chief Scientist of Empatica. She has a BS in Electrical Engineering from the Georgia Institute of Technology and an SM and ScD in Electrical Engineering and Computer Science from MIT. Picard is author of the book *Affective Computing*, and author or co-author of over two hundred and fifty peer-reviewed scientific articles. Picard's lab at MIT develops technologies to better understand, predict, and regulate emotion, including machine-learning based analytics that work with wearables and smartphones to help improve human mental and physical health and wellbeing.