Extraction and Interpretation of Deep Autoencoder-based Temporal Features from Wearables for Forecasting Personalized Mood, Health, and Stress

BONING LI, Rice University
AKANE SANO, Rice University

Continuous wearable sensor data in high resolution contain physiological and behavioral information that can be utilized to predict human health and wellbeing, establishing the foundation for developing early warning systems to eventually improve human health and wellbeing. We propose a deep neural network framework, the Locally Connected Long Short-Term Memory Denoising AutoEncoder (LC-LSTM-DAE), to automatically extract features from passively collected raw sensor data and perform personalized prediction of self-reported mood, health, and stress scores with high precision. We enabled personalized learning of features by finetuning the general representation model with participant-specific data. The framework was evaluated using wearable sensor data and wellbeing labels collected from college students (total 6391 days from N=239). Sensor data include skin temperature, skin conductance, and acceleration; wellbeing labels include self-reported mood, health and stress scored 0 – 100. Compared to the prediction performance based on hand-crafted features, the proposed framework achieved higher precision with a smaller number of features. We also provide statistical interpretation and visual explanation to the automatically learned features and the prediction models. Our results show the possibility of predicting self-reported mood, health, and stress accurately using an interpretable deep learning framework, ultimately for developing real-time health and wellbeing monitoring and intervention systems that can benefit various populations.

CCS Concepts: • **Representation learning → Wearable sensors**; *Recurrent autoencoders*; Personalized prediction; Network interpretability.

Additional Key Words and Phrases: health monitoring, neural networks, regression, stress, mood.

ACM Reference Format:
Boning Li and Akane Sano. 2020. Extraction and Interpretation of Deep Autoencoder-based Temporal Features from Wearables for Forecasting Personalized Mood, Health, and Stress. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 49 (June 2020), 26 pages. https://doi.org/10.1145/3397318

1 INTRODUCTION

Many physical and mental disorders manifest in various physiological and behavioral presentations before a diagnosis. For example, when patients with major depressive disorder first seek treatment, the presenting complaints can often be a recession of energy and happiness [64]. For Alzheimer's Disease, a physical decline in sleep and movement is observable to varying degrees in its earliest stages, prior to the presence of significant functional decline [62]. Such kinds of recession could have usually persisted for years before the patient eventually notice and go to the doctor, causing the optimal timing of treatment to be missed [45]. If daily health and wellbeing

Authors' addresses: Boning Li, boning.li@rice.edu, Department of Electrical and Computer Engineering, Rice University, 6100 Main St MS 366, Houston, Texas, 77005; Akane Sano, Department of Electrical and Computer Engineering, Rice University, 6100 Main St MS 366, Houston, Texas, 77005, akane.sano@rice.edu.

2474-9567/2020/6-ART49 $15.00
https://doi.org/10.1145/3397318

could be measured using ubiquitous sensors and assessed for early manifestation in an understandable way, clinicians would be able to provide early warnings and prevent severe disorders.

In recent years, the vigorous rise of both hardware and algorithms has opened up many opportunities to address this need. The industrial and commercial development of wearable sensors has enabled non-intrusive, ubiquitous, inexpensive, and continuous high-resolution data collection [7]. Moreover, deep machine learning methods such as convolutional neural networks (CNN) and recurrent neural networks (RNN) have shown outstanding performance on computer vision and natural language processing (NLP) tasks [84, 96**?**]. Consequently, recent studies have striven to transfer the methods, principles, and experience of deep learning to the relatively green fields of health sensing. For example, machine learning models have been developed to predict mood, stress, and health based on hand-crafted features computed from survey, weather, wearable, and mobile phone data [39, 39, 41, 50, 52, 57, 59, 60, 63, 76, 80, 85, 87, 87, 88, 91, 97–99, 106, 108, 109].

Nonetheless, engineering features from raw data requires domain expertise and human efforts. Later, selecting features for best fit can be time-consuming as well. To overcome these defects, many studies promoted deep learning to automatically learn features from raw data [4, 15, 28, 29, 44, 58, 66, 69]. However, a major drawback lies in the lack of clear explanation of those features and predictions. Until there is a way to accurately interpret the features learned and the predictions made, deep learning would hardly earn the trust of patients, clinicians, or other potential beneficiaries. Although some mechanisms exist to visualize the critical features or steps with slightly more intuitions (e.g., saliency maps and attention mechanism) [86, 107], it is still very difficult to associate those deep outcomes with a standardized and human-understandable description, especially for sequential sensor data within the health and wellbeing context.

In this paper, we propose an interpretable two-stage deep learning framework consisting of i) a recurrent autoencoder that learns to extract physiological and behavioral features from raw skin conductance (SC), skin temperature (ST) and acceleration (AC) data, and ii) a regression model that can predict personalized mood, health, and stress scores for any user based on her automatic features and subjective profile of personality and gender. We contribute novelty in the following three aspects,

(1) **Automatic learning of time-series representation from wearable sensors.** We introduce a hierarchical recurrent autoencoder to automatically learn efficient temporal features from wearable sensors. It can be finetuned using data of individual participants to incorporate personalized information and improve feature quality in general.

(2) **Adaptive personalized prediction of mood, health, and stress.** We leverage individual-based and cluster-based personalization strategies to predict wellbeing scores with higher precision than a generalized model. The cluster-based model is eligible to accommodate unseen users with high precision.

(3) **Interpretation of deeply learned features and predictions.** We provide interpretation on multiple levels. We leverage attention layers to visualize low-level time-step saliency, analyze correlations between auto-learned features and hand-crafted features for high-level explanations, and also discuss individual and clustered differences in coefficients of the prediction model.

Apparently, reliability, generalizability, and interpretability must be secured for any practical health and wellbeing prediction systems. We conclude that our method is reliable by showing that our auto-learned features outperform hand-crafted features in forecasting self-reported mood, health, and stress on a continuous scale. Secondly, we verify the generalizability of our framework by its reasonably good performance on predicting mood, health, and stress for new users using an adaptive personalization strategy. In addition, as a first attempt in the field of ubiquitous computing, we provide statistical interpretation and visual explanation of the learned features and personalized prediction models. Finally, we address computation, privacy, and potentials in implementing this framework as a real-time ubiquitous system.

2 RELATED WORK

2.1 Wellbeing Prediction Through Passive Sensing

Wellbeing refers to aspects of emotions, mood, and mental, and physical health [21]. Specifically, mood, health, and stress are three commonly studied wellbeing labels, and many studies demonstrated successful approaches to either detect or predict these labels using unobtrusively collected sensor data [39, 41, 50, 52, 57, 59, 60, 63, 76, 80, 85, 87, 88, 91, 97–99, 106, 108, 109]. For example, Zenonos et al. [108] presented HealthyOffice, a mobile application that can predict mood at work every two hours using past acceleration, temperature, and other wearable physiological data. More recently, Morshed et al. [63] focused on the early prediction of mood instabilities using activity, location, and audio data that were passively collected or computed from smartphones and wearables. Other researchers have targeted to infer various health conditions, such as Schizophrenia symptoms [97] and fatigue [88]. In [87], self-rated sick/healthy states were predicted by wrist-worn sensor and smartphone. Behavioral and physiological features (e.g., phone call duration, features related to the amplitude, shape, and rate of skin conductance responses) were computed and used to train a personalized machine learning model reaching 82.2% binary forecast accuracy. Additionally, passive sensor data could contain predictive information of stress [30, 35]. It has been shown that stress could be assessed by skin conductance [33], skin temperature [103], and acceleration [24]. Sano et al. [80] detected stress using a wrist-worn sensor of skin conductance and acceleration, as well as features of phone usage (e.g., calls, messages, screen on/off status). Closely related to our work, [106] predicted high/mid/low mood, stress and health labels for next day using 172 daily features such as number of steps, accelerometer weighted skin temperature, median skin conductance amplitude; the authors explored multiple machine learning algorithms, among which the highest accuracy was reported at 74%.

In general, predicting or forecasting wellbeing is typically more difficult than detecting or recognizing wellbeing, as Taylor et al. pointed out in [87], because of the latency. Since our ultimate goal is to aid in early intervention, to forecast fine-grained wellbeing status is what we should target at. More importantly, the aforementioned related studies entirely relied on hand-crafted features. While feature engineering can filter noise, reduce dimensionality, and combat overfitting, it requires domain expertise, usually subjective and suboptimal. As a result of convenient math formulas, crafted features may not adequately characterize the complicated patterns related to the outcome variables. We attempt to overcome these defects of hand-crafted features by leveraging data-driven deep learning methods to form feature extraction as an automatic learning process.

2.2 Deep Learning: Prediction, Representation, And Interpretation

In contrast to *crafting* features with fixed rules, autoencoders (AE), a type of artificial neural networks (ANN) trained for accurate reconstruction, may be leveraged to *learn* more complicated features from a large amount of raw data [47]. An ANN is based on layers of nodes and connections, which mimics the learning process of biological neurons or brains [70]. Depending on the architecture and dynamics, the basic concept of ANN can embody in many variations (e.g. multi-layer perceptron (MLP), CNN, RNN) for a wide range of real-world solutions including activity, emotion, and other wellbeing-related predictions [34, 68, 90]. Generally speaking, MLP is the "vanilla" form of ANN, commonly used on relatively simple datasets; CNN and RNN encompass more specificities, respectively in visual and sequential data [84, 96].

A single-layer AE with linear activation is equivalent to principal component analysis (PCA) which is essentially a linear transformation [15]. In most AE practice, however, non-linear activation functions are inserted to introduce non-linearities. Also, fully connected layers, convolutional layers, and recurrent layers can be stacked to make it capable of modelling complex functions. Consequently, the power comes with a cost such that PCA is faster and computationally cheaper than AE. Due to the high number of parameters, the good performance of AE relies on sufficient training data, appropriate regularization, and careful design [8].

Many studies have been carried out to explore the feasibility of distributing raw sensory data in AE so that wellbeing-related features can be automatically learned through back propagation in either unsupervised or supervised manner [29, 31]. For example, Mehrotra et al. [58] used an autoencoder to transform GPS mobility (displacement, change in displacement, and significant place) to features which performed better than hand-crafted features on predicting binary depressive states. Similarly, in essence, [66] integrated CNN and long short-term memory (LSTM) networks for human activity recognition based on 500 ms data segmentation from accelerometers and gyroscopes, and [44] also applied CNN-LSTM to recognize emotional levels using 45-minute recordings from smart-phones and wearables. On various tasks, these very recent studies have proven the performance advantage of avoiding manual feature engineering, revealing rising interest and promising potentials in tackling ubiquitous health and wellbeing prediction with deep learning principles and approaches.

An often attacked drawback of deep feature learning lies in the poor interpretability of its outcomes. Previously, efforts such as localization and visualization were made to investigate the black-box of image-recognizing CNNs [65, 105]. For sequential data, attention mechanisms can determine time-step saliency in RNN models, such as in image captioning [102] and steering angle prediction for self-driving cars [46]. Raghu et al. [72] proposed a singular vector canonical correlation analysis to probe interpretations of deep learning representation and dynamics. However, sensor signals are, by nature, less human-readable than image or text data. Even if we have generated accurate saliency heatmaps, it is still very difficult to find a standardized and intuitive way to explain to patients or doctors what those important features mean. Our study also recognized this challenge and provides a preliminary solution – we not only train deep representation and prediction models, but also associate the resulted features and predicting behaviors with interpretations using deep visualization and statistical approaches.

2.3 Personalization In Feature Extraction And Prediction

Exploiting individual differences in human physiology, behavior, and profile can positively boost one-size-fits-all models' performance. Multiple studies have confirmed an increase from 55% – 84% to 61% – 88% in binary classification accuracy by introducing personalized models to mood recognition [9, 92, 106, 108]. Unfortunately, personalization is often understudied with automatically extracted features due to the lack of labeled data in research or the lack of prior knowledge of incoming samples in practice [20, 25]. To overcome these defects, clustered multi-task personalization based on personal profile criteria can be an alternative with a compromise on prediction accuracy [113]. Utilizing knowledge of personality types, Ciocarlan et al. [18] assessed intentional engagement and explored persuasive message and activity interventions to improve wellbeing and prevent mental health problems. Closely related to our work, Taylor et al. [87] leveraged multi-task learning to forecast emotional wellbeing using features extracted from sensor, phone, and survey data with 78-82% classification accuracy.

However, none of existing works to our knowledge has investigated individual differences in feature extraction or representation learning. Even in personalization-focused studies, only the mapping from features to labels was ever personalized, yet from raw data to features was a unified process for all individuals. Motivated by the performance improvement brought by personalizing prediction models, we expect that personalizing the representation model may result in higher feature quality as well. Thus, in addition to a general representation model trained on all users' training data, we finetuned it to each individual to verify any further improvement of reconstruction accuracy or even prediction precision.

3 METHODS

3.1 Data Collection

Wearable and mobile technologies have assisted researchers to unobtrusively collect and monitor multiple body signals that can reveal one's internal state [82]. In this study, with the purpose of forecasting mood, health, and stress, we specifically focused on skin conductance (SC), skin temperature (ST) and acceleration (AC) signals

for the following reasons. SC is related with human physiological arousal, controlled by the sympathetic nervous system [10], and thus it can be an index of stress response [17]. ST can be reflective of many health problems (e.g. fever and heat exhaustion) [13, 73] and biological process (e.g. circadian rhythm and sleep) [48]. AC directly monitors human movement and sleep patterns, thus suitable for measuring energy and physical activities from which mental status could be inferred [59]. Another reason that we selected these three sensor modalities was to promote the idea of passive sensing. Also, these wearable physiology modalities could contain less privacy sensitive information than mobile phones, and thus it is a good point to start with. Our experiments, detailed in Section 4, showed that SC, ST, and AC data could produce robust and predictive measurements of future wellbeing.

The data were collected in the wild, using sensors worn on a wristband as one might wear a watch. The participants were 255 college students in New England. Data collection ran from 2013 to 2017, approximately from 30 days to 90 days from each student, in total 9189 days. We also collected daily wellbeing labels as the ground truth. Survey emails were sent to participants at 5 pm every day, asking for self-rated scores of mood (sad-happy), health (sick-healthy), and stress (stressed-calm) on a continuous slide bar scaled 0 – 100. Low scores indicated negative feelings and high scores were positive. The gender and Big Five Personality information were also available via standardized pre-study surveys [43]. Table 1 summaries our data, ground truth, and profile. The distribution of self-reported mood, health, and stress labels are biased to varied patterns and degrees, as displayed in Figure 1.

Table 1. Overview of the dataset

Category	Source	Type
Skin conductance (SC) Skin temperature (ST) Acceleration (AC)	Wrist sensor	Data (8 Hz)
Mood (sad-happy) Health (sick-healthy) Stress (stressed-calm)	Daily survey	Ground truth (0 – 100)
Gender Big Five Personality	Pre-study survey	Personal profile

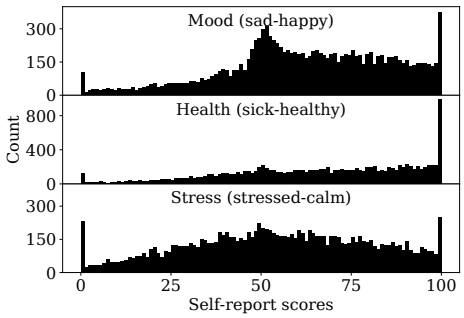


Fig. 1. Self-reported mood (sad-happy), health (sick-healthy), and stress (stressed-calm) scores.

The sensors sampled SC, ST, and AC data at a unified frequency of 8 Hz. Those data were preprocessed to remove artifacts and environmental noise. SC data were filtered using a 32nd FIR filter with a cutoff frequency at 0.4 Hz [14, 81]. A wavelet filter was adopted to remove artifacts from ST data [67], using Symlets 4 Scaling function and adaptive threshold by Stein's Unbiased Risk Estimate [111]. For the AC data, we computed the root sum squared to get acceleration magnitude. To test for multiple resolutions, we downsampled the data from 8 Hz to 4, 2, and 1 Hz via first order spline interpolation with Gaussian anti-aliasing [95].

We had missing data for various reasons (e.g., participant drop-out, sensor outage, survey incompletion, etc.). Given a participant and day, if the sensory missing rate was lower than 25%, the missing sensor data would be imputed with the corresponding channel means. Otherwise, or with survey responses missing, that day would be discarded from that participant's record. After data cleaning, 6391 days from 239 participants were valid. Finally, we applied intra-channel normalization to each participant and each day, where the daily SC, ST, and AC data were respectively normalized to range 0 – 1 in order to reduce bias and enhance robustness.

3.2 Feature Extraction

3.2.1 Baseline Feature Crafting. We computed 136 hand-crafted features from 24-hour SC, ST, and AC data according to previous work [87, 106]. For instance, SC data were characterized by the sum of area under the curve (AUC), median amplitude, count of peaks, etc.; ST features included min, max, and median values; AC features integrated step count, stillness percentage and mean movement step time. In addition, the impact of movement and temperature on physiology was considered as weighted SC features. These crafted features served to 1) produce the benchmark performance of personalization wellbeing prediction, and 2) provide vocabularies for interpreting automatically learned features.

3.2.2 Deep Feature Learning. Overview of Framework Architecture. We propose Locally Connected Long Short-Term Memory Denoising AutoEncoder (LC-LSTM-DAE) as the deep representation learning framework that learns and extracts physiological and behavioral features from raw wearable sensor data. We composed locally connected layers and LSTM layers with a denoising autoencoder to form a multi-modal deep representation learning model. In order for higher interpretability of auto-features, we decided not to mix the multi-modal data while the representation model was learning features. We trained independent representation models for each channel and concatenated single-channel features to form the final feature vectors. Twenty-four-hour sequences of SC, ST, or AC data were sent to the model as inputs, divided into eight 3-hour frames to address the temporal characteristics of sensor data. In each time frame, we adopted the same static MLP structures. Outputs of eight time frames were later joined sequentially by an LSTM encoder, and the final hidden state of the bottleneck layer was extracted as the learned features. It was then added Gaussian noise and copied to the input of the symmetric LSTM decoder, unrolled to the former dimensions, and then decompressed to the input dimensions via a stack of linear layers symmetric to the static encoder (Figure 2). Each component will be detailed in the following paragraphs.

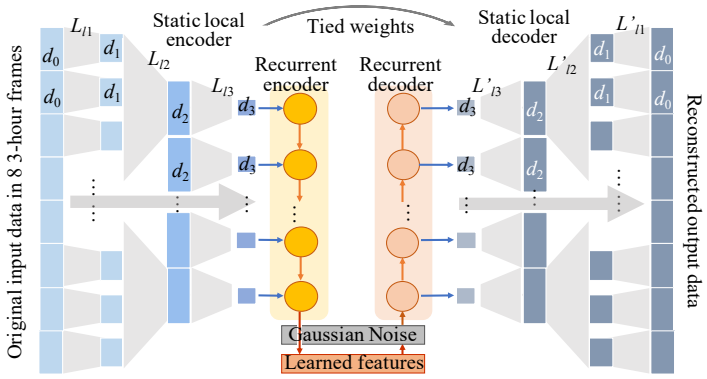


Fig. 2. Schematics of LC-LSTM-DAE for automatic feature extraction from raw sensor data.

Locally Connected Layers. For each separate time frame, 3-hour single-channel raw data in 8 Hz (3 hr × 60 min × 60 sec × 8 Hz = 86400 input nodes) would already cost over a billion trainable parameters for a plain neural network, which could lead to overfitting or difficulty to converge. To avoid overparameterization, we adapted hierarchical local connections as Equation 1 describes.

$$a_j^l = \sigma\Big(\sum_{k \in F_i^l} w_{jk}^l a_k^{l-1} + b_j^l\Big) \tag{1}$$

where in layer l, a_j^l denotes the j-th node; F_i^l is the index set of nodes belonging to time frame i; w_* and b_* are corresponding trainable parameters (weight and bias). The Rectifier Linear Unit (ReLU) is used as the activation function $\sigma(\cdot)$.

For each channel, LC-LSTM-DAE starts with a locally connected layer which is densely time-separated. In the first layer, we divide the 24-hour raw data input into 1440 1-minute non-overlapping windows, each containing d_0 locally connected nodes. Parameters within each 1-minute window in Layer L_{l1} are updated independently and do not interact with other windows. Within each window, the local input is projected into a d_1-dimensional space, resulting in a total of $1440d_1$ nodes at the input of L_{l2}. Terminologically, we say that the first local layer L_{l1} has "one-minute temporal windows". Starting at L_{l2}, we broaden the temporal windows to 3 hours (8 windows per day: 0 – 3 am, 3 – 6 am, 6 – 9 am, 9 – 12 pm, 12 – 3 pm, 3 – 6 pm, 6 – 9 pm, 9 pm – 12 am). Dimensionality reduction from $1440d_1 \text{—}[L_{l2}]\rightarrow 8d_2 \text{—}[L_{l3}]\rightarrow 8d_3$ is completed in multiple layers with ReLU activation stacked sequentially. The final dimensionality of the learned features is $8d_3$.

Static Autoencoder. Autoencoders are essentially unsupervised neural networks, originally developed to learn efficient data representation, typically for anomaly detection [77], smart imputation [71] and nonlinear dimensionality reduction [94]. Starting with the original data at the input, the autoencoder encodes it to a lower dimensional feature space, followed by a reconstructing decoder that tries to generate a signal from the reduced representation. The autoencoder is trained in mini-batches through multiple iterations such that the generated signal should be as close as possible to its original input.

Architecturally, the simplest form of an autoencoder is a feed-forward, non-recurrent multilayer perceptron whose output layer has the same number of nodes as the input layer. The objective function is defined as some distance between the output and the input. Encoding is preserved at the bottleneck layer. To improve the robustness and richness of the condensed information, various techniques exist beyond the basic encoder-decoder architecture. For instance, variational autoencoder (VAE) and denoising autoencoder (DAE) are most common in practice. VAE draws representation from distribution, so the extracted features will have randomness. To ensure reproducibility, we adopt DAE by adding Gaussian noise $\sim \mathcal{N}(0, 0.1)$ to the representations, followed by a decoder forming a symmetric architecture as the encoder.

Besides, autoencoders with tied weights have important advantages including i) less parameters to learn, ii) more geometrically adequate coding, and iii) tied wights can act similarly as regularization. Equation 2 gives the form of a single-layer tied weight autoencoder,

$$f_\theta(x) = \sigma_2(b_2 + W_{dec}\sigma_1(b_1 + W_{enc}x)), \text{where } W_{dec} = W_{enc}^{\mathrm{T}}. \tag{2}$$

In Equation 2, b_1 and b_2 respectively denote the bias terms of encoder and decoder; σ_1 and σ_2 are activation functions; W_{enc} is the trainable weights of encoder, while W_{dec} is the decoder weights which are not trainable and derived from W_{enc} as its transpose. With linear activation functions as σ_1 and σ_2, it would be equivalent to the principle component analysis (PCA). Our proposed model is a stacked version with multiple non-linear activated layers connected sequentially. The objective function is based on mean square error (MSE) and L2-norm regularization, shown as follows,

$$\hat{\theta} = \text{argmin}_\theta \frac{1}{2m}\Big[\sum_{i=1}^{m}(f_\theta(x^{(i)}) - x^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2\Big]. \tag{3}$$

where θ denotes the to-be-optimized autoencoder parameters, and $f_\theta(\cdot)$ is the encoding-decoding transformation; input raw data x is of length m; λ is a constant coefficient that introduces the L2-norm regularizer.

Recurrent Autoencoder. Recurrent neural network and its variants have been proved a success in multiple human related applications including sleep classification, activity recognition and mental stress level prediction

[32, 79, 93]. When incorporated in autoencoders, recurrent networks can summarise information from sequential data, which agrees with the nature of continuously collected sensor data. Therefore, we inserted a 2-layer LSTM as the core encoder-decoder component in LC-LSTM-DAE, fed on 3-hour intermediate static features at each time step. Its form can be simplified as follows,

$$\begin{aligned} \text{Encoder: } h_t &= f_\phi(Uh_{t-1} + Wf_{\text{static}}(x_t)) \\ \text{Decoder: } h_t &= f_\psi(Uh_{t-1})) \end{aligned} \tag{4}$$

where h_t is the hidden state at time step t; W is the input-hidden weight matrix and U is the recurrent weight matrix. f_ϕ, f_ψ, and f_{static} abstract other internal transforms (gates mechanism, activation functions, stacked layers, etc.) in LSTM encoder, LSTM decoder and static encoder.

3.3 Wellbeing Prediction

3.3.1 Generalized Model -- One-Size-Fits-All LSTM With Recurrent Batch Normalization. Long short-term memory (LSTM) is a recurrent ANN that is well-suited for processing time series where there can be lags of unknown duration between critical moments. A growing number of studies have successfully achieved and promoted prediction of emotion, stress, and other wellbeing labels with LSTM-based approaches using speech, accelerometer, and other modalities [3, 93, 106].

We built a 2-layer stacked LSTM followed by a single dense layer fed on multi-day auto-learned features, one day per time step. Dropout rate was set at 0.3 in all LSTM layers and 0.5 in the dense layer. Sequence-wise batch normalization was also applied, as suggested for a speech recognition problem where Amodei et al. [5] demonstrated that sequence-wise batch normalization in RNNs substantially improved both final generalization error and the speed of convergence. The insertion form is given in Equation 5,

$$h_t^l = \sigma\left(\eta \frac{W^l h_t^{l-1} - \mathrm{E}[W^l h_t^{l-1}]}{\sqrt{\mathrm{Var}[W^l h_t^{l-1}] + \epsilon}} + \beta + U^l h_{t-1}^l\right). \tag{5}$$

where $\mathrm{E}[\cdot]$ and $\mathrm{Var}[\cdot]$ are the empirical mean and variance over a single time-step $(t-1)$ of a minibatch. The learnable parameters η and β respectively scale and shift each hidden unit in layer l. The small positive constant ϵ is included for numerical stability. The sequential dependence between time-steps prevents averaging over all time-steps.

3.3.2 Personalized Models -- Multi-Task Linear Regularized Models. Multi-tasking learning (MTL) algorithm can optimize multiple different yet related tasks together [11] by sharing some information across tasks in the learning process. The final generalization effect of MTL is usually superior to that of one-size-fits-all learning (no task-specific information) and single-task learning (no shared information) [53].

One fundamental decision to make for MTL is the definition of *tasks*. Intuitively, one would desire intra-task data to be similar while inter-task data different, exactly like our study participants who shared similarities yet also exhibited differences. They were recruited on campus and many of them took same classes or knew each other, thus they might share commonalities in sleep, exercise, curriculum, and activity patterns. In this paper, we compared between MTL predictions with *individual participant* as tasks and with *clusters of participants* as tasks.

Alternatively, the individual-as-task MTL can be viewed as a special case of the clusters-as-task MTL. To create participant clusters, K-prototypes clustering was applied to their profile surveys (gender and Big Five Personality). K-prototypes was firstly proposed in [37] for clustering datasets with mixed numeric and categorical values. The number of clusters was determined at the highest mean Silhouette score of all samples [75]. The best value of Silhouette coefficient is 1 and the worst value is -1. A negative value generally indicates that the sample is assigned to a wrong cluster.

Linear models are interpretable and widely-used in solving regression problems. We adopt $\ell_{2,1}$ regularized MTL linear regression model whose objective function is as follows,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} ||\theta_i^T X_i - Y_i||_2^2 + \lambda_{2,1}||\theta||_{2,1} + \lambda_2||\theta||_2^2 \tag{6}$$

where X_i represents the input matrix of the i-th task, and Y_i is the label of samples belonging to that task; θ is the weight matrix. λ_2 controls the ℓ_2-norm penalty; $\lambda_{2,1}$ controls the $\ell_{2,1}$-norm penalty.

3.4 Interpretation

To provide actionable and effective interventions, which is critical to make this system real-world beneficial, we emphasize the interpretability of our system. That said, we need to not only develop algorithms that can provide good wellbeing forecast but also design mechanisms to deliver human understandable interpretation to those features and forecast. Good interpretation of results can convince patients and doctors to benefit from our solution.

In this study, we provided interpretation at multiple levels. First, we employed attention mechanisms in multi-day prediction models to give insights to the rise or decay of sensing data impact on the users' wellbeing states. In addition, we analyzed a correlation between the learned features and the crafted features to broaden our understanding of the former. Additionally, we performed K-means clustering to the MTL weights to reveal the relationship between individual differences and prediction performance.

3.4.1 Attention Mechanism. Bahdanau et al. [6] proposed the attention mechanism to allow a recurrent decoder to attend to different parts of a long input sequence at each step. Ever since, attention has been applied in many NLP problems vastly beyond encoding-decoding models. In many-to-one classification or regression models, the idea is to take the importance of every time-step from the inputs into consideration, as shown in Figure 3 [101]. The softmax-normalized attention weights quantify the importance scores of a sequence.

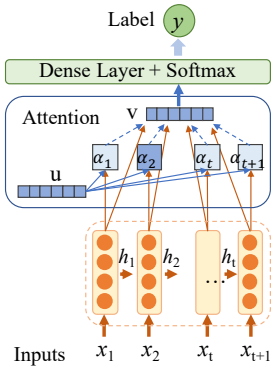


Fig. 3. Illustration of attention mechanism in many-to-one LSTM network, where h_t denotes the hidden state at time-step t; is the attention weight; u and v are the importance vector and the context vector.

3.4.2 Correlation Analysis. Correlation analysis is used to evaluate the strength of a relationship between two variables under certain assumptions [27]. In this study, Pearson correlation was used as a preliminary probe of the "physical meaning" of the auto-learned features. By correlating them with predefined hand-crafted features, we desired to establish if there existed significant correlations between learned and crafted features of wearable sensor signals. Examining the top-5 learned-crafted correlations, we might be able to translate what the representation model had learned into human understandable language.

3.4.3 MTL Weight Analysis. In the personalized prediction, to identify the critical features, we looked for features that i) contributed the most in a task or ii) resulted in diverse weight coefficients across different tasks. The first target can be simply achieved by searching all task-specific weights and identifying the highest contributing features. For the second target, in particular, we clustered the weight vectors of all 239 individual-based tasks using K-means and Silhouette score evaluation, similar to the principle described in 3.3.2.

4 EXPERIMENT

Our goal is to use deep learning methods to automatically learn and extract efficient physiological and behavioral features from high-resolution wearable sensor data. Furthermore, we show by fitting a wellbeing prediction model that the auto-learned features can be used to predict multiple wellbeing labels on a continuous scale from 0

– 100 with high precision. The evaluation metrics are reconstructive mean square error (MSE) for representation models and mean absolute error (MAE) for prediction models.

Analysis of variance (ANOVA) [38] test was used in testing differences among more than two groups of data. Following ANOVA, Tukey HSD test [1] was used for testing the difference between each pair of two groups across the whole data set. Paired t-test was used for comparing the performance between the models using different features sets (e.g., crafted features, static auto features, temporal auto features, etc.).

The experiments were conducted on a Ubuntu 18.04.1 LTS Linux machine. Models were trained with dual NVIDIA GeForce RTX 2080 Ti GPUs. Representation models and LSTM prediction models were implemented using the deep learning platform PyTorch 1.0. Linear MTL prediction models were adapted from the MALSAR toolbox [114].

4.1 Representation Learning

Training a deep learning model is usually time-consuming and computationally expensive. In the meantime, many deep learning methods in health applications are limited by the training set size. Multiple recent studies [2, 16, 56] supported "less is more" – reducing training set size could be harmless to the trained model; doing it smartly could even give the model better generalization ability and robustness. Coming to our study, we were also interested in the minimum acceptable size of training set for time/energy saving purpose.

The representation learning model was optimized via the Adam algorithm, with the learning rate being 0.003 and trained for 100 epochs. β_1 and β_2 were set 0.9 and 0.999 respectively, and the weight decay was $1e^{-6}$.

Cross validation was conducted in 4 folds with the train/validation/test ratio as 60%/20%/20%. Section 4.1.5 details two split schemes differentiated by user dependency. In brief, under user-dependent settings, the training set contains data from all users; i.e. after training completes, the model is guaranteed that no unseen users would be presented as a test sample. On the contrary, the user-independent case requires the model to handle both seen and unseen users at the test stage.

A prior work [51] showed that the dimensionality of auto-learned features did not have as a significant impact on the final prediction performance as input resolution or hyper-parameter configuration. Therefore, an empirical final dimensionality of 48 was chosen for this study.

4.1.1 Input Resolution. We compared the reconstruction loss (MSE) using 8, 4, 2, and 1 Hz raw sensor data input to the representation learning model. Lower-resolution data were downsampled from the original 8 Hz using first-order spline interpolation. Gaussian smoothing was performed to avoid aliasing artifacts.

4.1.2 Amount Of Training Data. We probed into the impact of training data amount on the reconstruction loss for LC-LSTM-DAE by randomly selecting different sized subsets for training. The baseline amount was set to be 60% of the entire dataset, for approximately 3900 days in total, or 18 days per participant. In addition, we fixed validation and test sets and tested 2/3 (40% of the entire dataset size), 1/3 (20%), and 1/6 (10%) sized training sets.

4.1.3 Dynamic Vs Static Architecture. We investigated the effect of temporal information incorporated by the LSTM encoder-decoder in the representation model by comparing validation losses with or without LSTM layers wrapping over the bottleneck layer. Without the LSTM encoder-decoder, the final features would simply be the concatenation of the output at L_{l3} (Figure 2).

4.1.4 Personalized Representation Learning. Another experiment was carried out where we introduced individual differences to the learning of temporal features. More specifically, this preliminary experiment concentrated on the LC-LSTM-DAE model under the user-dependent assumption. The trained general LC-LSTM-DAE was finetuned using each participant's data for a maximum 100 epochs. When finetuning for a participant, if a new lowest MSE was not reached for more than 30 epochs, the finetuning for this participant would automatically

terminate. All other learning parameters were kept the same as they had been when the general LC-LSTM-DAE model was trained. We hope that this experiment, as a first attempt to personalize the mapping from raw data to features, would address the potentials of low-level data personalization.

4.1.5 Data Preparation For User-Dependent And User-Independent Experiments. Two schemes were designed to split 60% train, 20% validation, and 20% test data for user-dependent and user-independent experiments.

In the user-dependent setting, the training set should contain data from every user to ensure that any users in validation and test sets have data also in the training set. While shuffling data, the unique user-date identifier was kept in track to construct cohorts for MTL. Also, we designed the split to accommodate for 4-fold cross validation. In specific, we iterated over all users; for each user, we got all of her days in the dataset, shuffled those days, evenly split into 5 sub-lists, randomly chose and left out one as the test set, and then the rest 4 sub-lists would act as the validation set in turn.

The user-independent experiment was designed to simulate the scenario that we would face when promoting this system to larger-scale tests and real-world usage. Validation and test sets were composed of users that were not in the training set. Although it would be more challenging for the model to learn and generalize under such setting, constructing training sets was actually much simpler – just by taking all users and performing a random split; each user should always take all of her days to one of the train, validation or test set.

4.2 Wellbeing Prediction

When regularizing the MTL wellbeing prediction model using auto-learned features, tuning the penalty coefficients can be critical. Otherwise, it could cause the loss of dimension because the number of features has become relatively small. The constants λ_2 and $\lambda_{2,1}$ were determined via grid search as 0.15 and 0.1.

4.2.1 Auto-Learned Vs Hand-Crafted Features. We further compared two sets of wellbeing prediction performance. One was using a set of 48 auto features learned and extracted by our deep representation model, and the other was based on 136 crafted features that provided benchmark results.

4.2.2 Number Of Previous Days. Previous studies [39, 106] showed that using hand-crafted features, exposing to longer past could sufficiently improve the mood and health classification performance. In this study, although the input to prediction models has been changed to auto-learned features and the problem domain was set to regression, we would still expect the gist to hold true. By incorporating more information from the past, the performance should be improved. To validate this hypothesis, we tested 1 and 7 days of features for wellbeing prediction.

4.2.3 Personalization. Depending on the user-dependency policy used in producing the auto-learned features as described in Section 4.1.5, the personalization approach for wellbeing prediction should be adapted accordingly. Either individual-based or cluster-based strategy was applied. In terms of the cluster-based strategy, we applied K-prototypes to cluster mixed user profile data in order to enable inference for unseen users. We adopted Silhouette score to find the optimal number of user groups. In the user-independent case, the clusters were obtained on a training set of 147 users. The Silhouette scores were then derived by assigning all 239 users' data to the trained clusters. We examined the mean values of the scores from 10 random trials, and higher scores were desirable.

5 RESULTS

5.1 High Vs Low Input Resolution For Learning Auto-Features

Figure 4 shows, during the training of the representation model, the validation MSE vs trained epochs with different input data modalities and resolutions. Because different sensors performed differently, we display them in separate plots. Overall, it can be observed that the highest resolution of 8 Hz always produced the lowest MSE for all sensors. More specifically, SC was not very sensitive to resolutions greater than or equal to 2 Hz; ST did not seem

to be sensitive to resolution at all levels, in that the drop in training curves was almost identical; AC concluded at a roughly logarithmic decrease with resolution degradation in the representation learning performance.

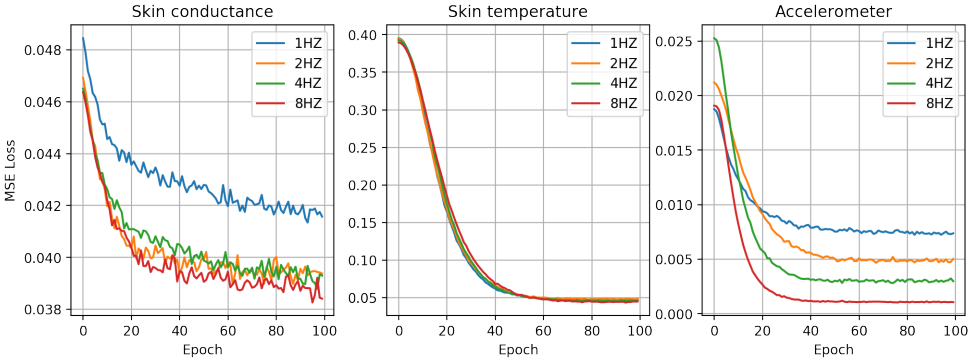


Fig. 4. Reconstruction loss on validation set with different input data resolution for each channel.

5.2 Amount Of Training Data For Learning Auto-Features

Figure 5 displays four user-dependent validation curves of reconstruction loss with different amount of training data for the representation model. It obviously shows that adequate and diverse training data can help the model converge faster. Nonetheless, with a limited amount of training data, the unsupervised model could still eventually converge to a stable state such that the original data could be reconstructed equally well from the learned features, given enough times of mini-batch iterations. These curves were all channels combined, because three channels behaved very similarly. Therefore, we can safely set the focus on the comparison of the training set size itself.

5.3 Temporal Vs Static Autoencoder Architecture

Figure 6 compares the curves of reconstruction loss on the user-dependent validation set with different AE architectures as described in Section 4.1.3. It emphasizes the advantage of the temporal component – the LSTM encoder-decoder wrapper – in LC-LSTM-DAE. Although both models started off similarly, without the temporal component, the static AE could not reach a local optimum as far as the temporal AE. Here we again combined channel-wise losses because all channels shared a consistent tendency.

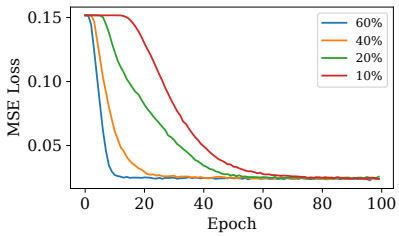


Fig. 5. Reconstruction loss vs training-set size.

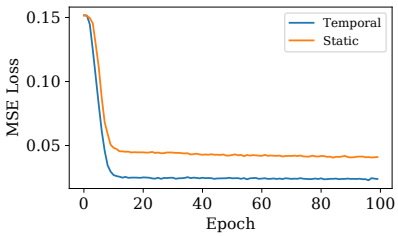


Fig. 6. Reconstruction loss vs AE architecture.

5.4 Personalization In User-Dependent Representation Learning

Figure 7 shows the histogram of changes in participant-specific MSE by finetuning the general LC-LSTM-DAE to each individual. Given a participant, a negative change in MSE indicates a drop in reconstruction errors and is thus desirable; otherwise, we would recognize the personalization as failed for this participant. In this experiment,

we observed that the personalization was favorable for a majority of 76% participants. In terms of the failure cases, possible reasons may include overfitting, simpler underlying structure of data or the limited amount of finetuning samples.

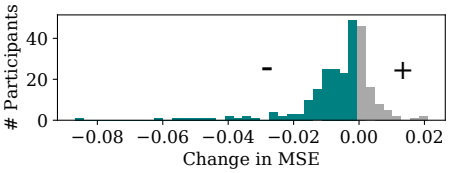


Fig. 7. Histogram of user-specific MSE changes introduced by finetuning the general representation model to each individual participant (−: benefited; + worse-off). 76% participants actually saw a reduction in reconstruction loss, i.e. an increase in feature quality.

5.5 Auto-Learned Vs Hand-Crafted Features: Prediction Performance Comparison

To demonstrate that LC-LSTM-DAE can learn efficient features to predict multiple wellbeing labels, in Table 2 we compared its performance with benchmark results derived from prior works on the same topic [51, 106]. The prediction model was individual-as-task MTL, under the user-dependent settings. The comparison results include i) crafted vs static vs temporal features, ii) 1-day vs 7-day temporal features, and iii) non-personalized vs personalized temporal features.

To begin with, we found that for any labels in Table 2, the temporal features always demonstrated significantly higher precision than the static features and crafted features ($p < 0.05$). Then, with regard to temporal features, 7-day concatenation significantly outperformed 1-day features in predicting mood and health ($p < 0.05$), but not in stress. Interestingly, personalizing the mapping from raw sensor data to features did not seem to accomplish much improvement in wellbeing prediction. Based on the LC-LSTM-DAE, only 1-day prediction of mood was significantly improved by adopting personalized features, suggesting that instant feelings of happiness/sadness could be a highly personalized process from physiology to self-perception.

Table 2. Wellbeing prediction mean absolute errors (MAE) using individual-as-task MTL on different feature sets (Mean±S.D.).

Label	**Crafted Features**	**Auto Features**				
		Static (LC-DAE)	Temporal (LC-LSTM-DAE)			
		1-day			7-day	
(PRS*)	-	-	-	+	-	+
Mood	16.4±0.3	14.6±0.3	14.3±0.2	14.0±0.3	14.1±0.2	**14.0±0.2**
Health	15.8±0.3	14.3±0.5	12.6±0.3	12.7±0.4	**12.4±0.3**	12.5±0.2
Stress	16.7±0.3	15.7±0.3	15.0±0.3	15.0±0.6	**15.0±0.3**	15.0±0.7

∗ PRS denotes whether or not the corresponding features were personalized.
−: Non-personalized features (same extraction rule for all participants);
+: Personalized features (different extraction rules across participants).

5.6 Generalized Multi-Day LSTM Wellbeing Models

Figure 8 compares the one-size-fits-all LSTM wellbeing prediction performance using multi-day static and temporal features. Significance indicators are denoted beside wellbeing labels to indicate whether ANOVA test rejects null hypothesis on the difference between the corresponding static and temporal feature performance. Over all labels and features, 7-day prediction was significantly more precise than 1-day prediction. Nevertheless, we found that temporal features showed a significant advantage over static features only on mood and stress using 7-day data.

Figure 9 shows the distribution of attention weights (sum to one) given to each time-step in a 7-day prediction model. Obviously, the prediction model believed that features from the nearest day was the most influential on future wellbeing, and the saliency gradually decreased as we moved away from the current time point. This could indicate that people's current state of wellbeing is likely to be affected by things that occurred several days ago.

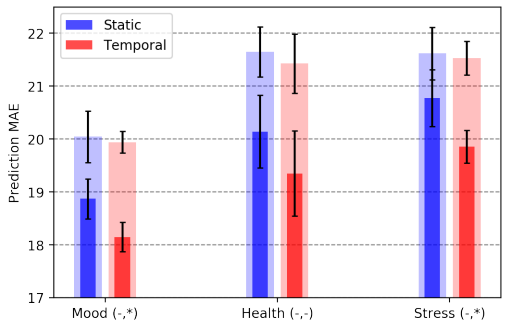


Fig. 8. Generalized wellbeing models' prediction performance using multi-day static and temporal features. Inner bars are 7-day results, and outer bars are 1-day results.

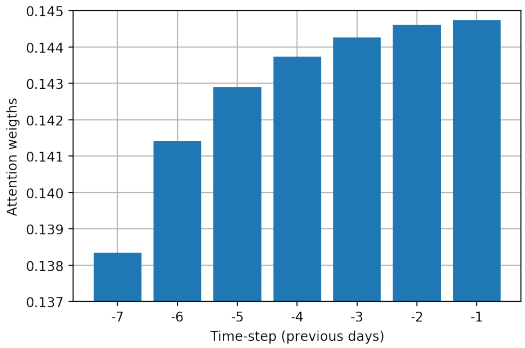


Fig. 9. Attention weights distribution on time steps in LSTM prediction model using 7-day temporal features.

5.7 Clustering Based On Profile

Figure 10 illustrates that following an initial drop around the # of groups=2-12, the Silhouette scores constantly increased with #groups and reached the highest 0.60 at the # of groups=147, when each user in the training set became a unique cluster centroid. In this case, a new user would be assigned as to "be like" one of the existing users to whom she has the closest gender and personality profile.

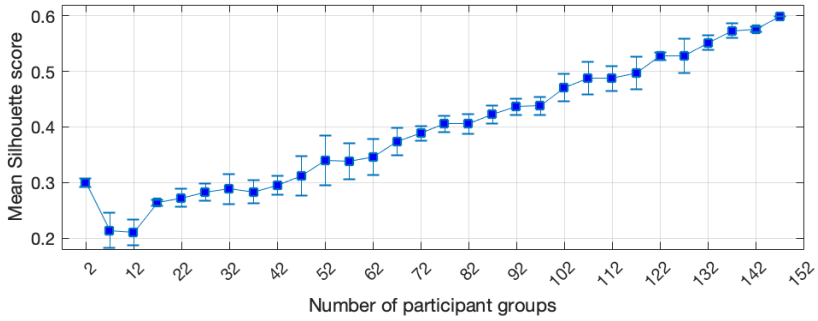
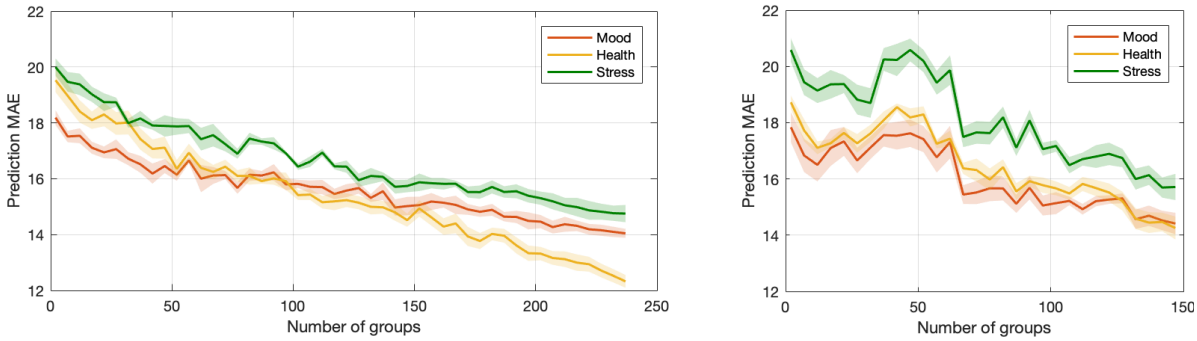


Fig. 10. Silhouette score vs number of user groups.

5.8 User Dependent Vs Independent Policies

Figure 11 shows an overall tendency of decreasing MAE with the increasing number of user groups, under both user-dependent and independent settings. The best performance of user-dependent MTL prediction was superior to the best of user-independent MTL models, which was expected because i) in the autoencoder, failure to see all users' distribution of raw sensor data might create some implicit difficulty for the generalization of auto features; ii) more importantly, being able to see all users beforehand became a big advantage in the explicitly personalized MTL prediction model.



(a) User-dependent (b) User-independent

Fig. 11. Wellbeing prediction MAE vs number of user groups.

In the user-dependent case (Figure 11a), the prediction performance steadily grew as the number of user groups grew. Eventually, the best performance was reached with 239 groups (equal to the number of users in the training set), which made cluster-based personalization equivalent to individual-based personalization, given that no new users would be introduced to the trained model. With this setting, the prediction MAE values for mood, health, and stress were respectively 14.1±0.2, 12.4±0.3, and 15.0±0.3.

In the user-independent case (Figure 11b), with # of groups=147 (equal to the number of users in the training set), the prediction performance reached its peak. In other words, the best MTL strategy appeared to be the following, i) keep individual-based tasks in the training set; ii) for a new user, find her closest existing user and "pretend" that they are the same person; iii) then the trained MTL weights can naturally be applied to the unseen test user. This scheme was the same as what we derived from Section 5.7. The lowest achieved prediction MAE values were 14.5±0.4 for mood, 14.4±0.4 for health, and 15.7±0.5 for stress.

5.9 Interpretation Of Features And Predictions

5.9.1 Correlation Analysis. We computed the correlation matrix between the corresponding modalities of auto-learned features and hand-crafted features. The significance-filtered ($p < 2.3\times10^{-5}$, the adjusted p-values [36]) correlation heatmap for daily features is shown in Figure 12. Generally speaking, from what we can observe, the temporal representation model seemed to have paid more attention to evening physiology (17H+:SC, 17H+:ST) as well as night-time activities (0-3H:AC, 17H+:AC).

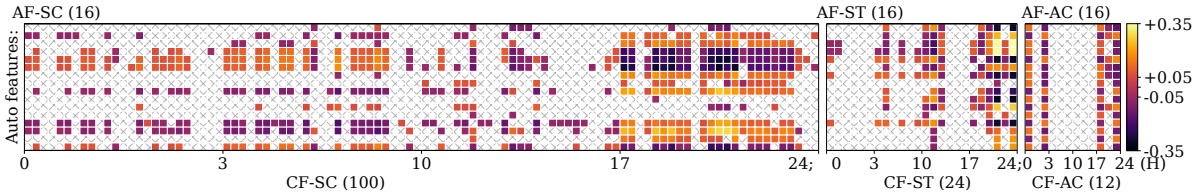


Fig. 12. Correlation matrix between auto-learned features (AF-) and hand-crafted features (CF-) pairs. The dimensionality corresponding to each modality/type of features were annotated in parenthesis. Annotated on the color bar is the adjusted significance threshold of ±0.05 ($p< 2.3\times10^{-5}$). The auto SC, ST and AC features were independently learned from the corresponding modalities and were not personalized.

In Table 3, we identify the best interpreted auto features, one for each channel, defined by the highest total correlations given by crafted features. Correspondingly, we list the names of top-5 crafted features with the highest absolute correlations. In general, physiological features (SC, ST) were more interpretable than activity

features (AC). This could be caused by the fact that we had a smaller corpus (i.e. less crafted features) for activities than physiology.

Table 3. Most interpreted auto features for each sensor modality, as well as the top-5 highest correlated crafted features with associated correlation coefficients ($p < 2.3 \times 10^{-5}$).

Auto Features	**Crafted Features**				
	top-1	top-2	top-3	top-4	top-5
AF-SC#6	17H+: unnormalized mean -0.30	17H+: normalized s.d. -0.30	17H+: count of peaks -0.29	17H+: 30-min median peaks -0.28	17H+: area under curve (AUC) -0.28
AF-ST#5	17H+: median while still -0.31	17H+: raw value median -0.30	10H-17H: raw value min 0.18	17H+: raw value s.d. -0.15	3H-10H: raw value min 0.11
AF-AC#5	17H+: stillness percent 0.16	17H+: step count -0.13	0H-3H: stillness percent 0.13	0H-3H: step count -0.12	17H+: mean movement step time 0.08

5.9.2 Weight Analysis. According to Table 3, health was the best predicted wellbeing label. Thus, we present weight analysis based on the health-predicting individual-as-task MTL model. To discover inter-personal similarities and differences, we quantified the MTL weight patterns by clustering the coefficients and looking for significant inter-cluster differences. The clustering achieved the highest Silhouette score of 0.59 at 2 clusters of 125 and 114 participants respectively. Then we computed intra-cluster mean values and tested for the significant difference via ANOVA. We found that 27 out of 48 auto features (#SC=12, #ST=7, #AC=8) were significantly different ($p<0.05$) between two clusters. We also found that SC produced the dominant inter-cluster differences, not only in terms of the number but also with regard to the significant gaps. The top-5 biggest gaps in two clusters' mean coefficients were all produced by SC features, namely AF-SC#12, #4, and #14, ranging from 11.6 to 3.72. For the ST and AC features, the most different-between-clusters features were AF-ST#8 and SF-AC#4 with the gaps being 0.58 and 0.92 respectively. Moreover, we confirmed that AF-SC#12 was a critical feature, because it was not only significantly different between two clusters but also opposite in sign. It could indicate that different people could react in different directions toward this feature. Such features are particularly of our interest because eventually, we would have to carefully consider controversial interventions regarding the modifiable behaviors related to such features, as it may lead to not only weak but even opposite responses.

Interestingly, we found that AF-SC#12 had significant correlations only with 10H-17H: SC unnormalized median (-0.06) and 10H-17H: SC median peaks in 30 min (-0.06) – the feature that was critical to personalization was not among the best-interpreted auto-features. A possible explanation could be that the representation model has learned beyond human knowledge. However, we cannot yet claim that AF-SC#12 was a strong predictor of health. We can only safely state that it was a highly individual-dependent feature, thus likely to be critical to personalized prediction and intervention.

To take one step further, we inspected the differences on an individual level. In Figure 13, we visualize the weight coefficient vectors of three example participants P1, P2, and P3. We intentionally chose them with health-prediction performance distributed at the 1%, 20%, and 40% percentiles (1.4±0.5, 5.8±1.6, and 9.1±1.3). By comparing task-specific weights of varied performance, we could get some insights into how the prediction of health was delivered in different individuals. For instance, it can be observed that the best-performing P1 had more stable weights among three modalities compared to P2 and P3. Some AC features of P2 seemed to be emphasized while AF-AC#5 noticeably stood out as it contributed strongly in the negative direction. The highlight in P3 was AF-SC#12 being negative with a large magnitude. The three participants showed a similar trend in treating ST features, although the ST coefficients of P3 were constantly lower than the others.

We checked the interpretations of these features given by the crafted ones, and we found that AF-SC#4 could be best interpreted as being negatively correlated with 17H+: SC 30-min median peaks (-0.23) whereas AF-SC#14 had a strong positive correlation with 17H+: SC normalized median (0.28). Besides, AF-ST#8 was weakly correlated with 17H+: ST raw value median (-0.06), while SF-AC#4 mostly correlated with 0H-3H: AC stillness percent (-0.13). AF-AC#5, the feature in which P3 found interest, was moderately correlated with 17H+: AC stillness percent (0.16).

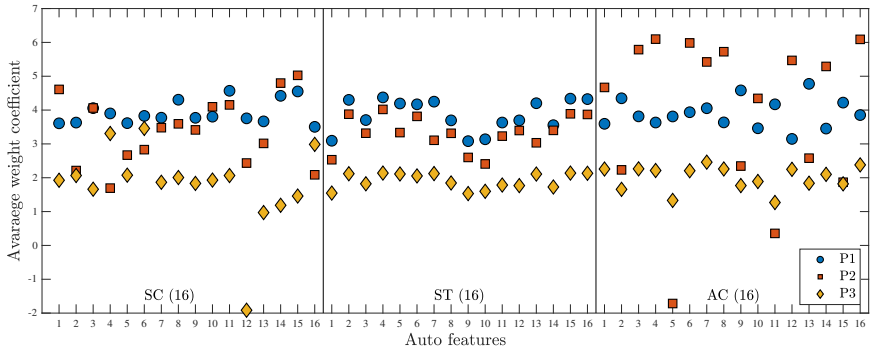


Fig. 13. Task-specific weights vectors for 3 example participants (P1, P2 and P3) whose prediction performance were 1.4±0.5, 5.8±1.6 and 9.1±1.3, respectively within the 1%, 20% and 40% percentiles.

6 DISCUSSION

As part of our analysis, we found that our proposed recurrent autoencoder framework LC-LSTM-DAE could efficiently learn temporal features from high-resolution sensor data, and those features were reliable estimators for forecasting one's subjective wellbeing. We highlight three major observations: i) temporal autoencoders can learn features that are more informative than static autoencoders; ii) subjective wellbeing can be forecast using auto features learned from passive sensing data in personalized manners; and iii) automatically learned physiology (ST and SC) features were more interpretable than behavioral (AC) features, as we found more significant correlations in the former with crafted features. In this section, we compare our results with prior work and outline the implications of our results for researchers in ubiquitous computing community who are interested in promoting human wellbeing with deep learning solutions.

6.1 From Features To Wellbeing Labels

For the representation model, we tested different configurations. First, we observed from Section 5.1 that different modalities responded differently to changes in input dimension or resolution. We further observed that modal sensitivity to the resolution had a positive correlation with its dominant frequency. In our case, although the sampling rate was unified, the nature of collected sensor data and different denoising process could introduce varied frequencies (AC > SC > ST). The observation that higher resolution led to lower reconstruction loss provided evidence that LC-LSTM-DAE was able to learn features efficiently from high-resolution data. In practice, we need to balance between learning performance and computational cost, as both can be increased by higher resolution.

In general, features that can achieve lower reconstruction loss are desirable. However, does lower reconstruction loss necessarily lead to higher prediction accuracy? Based on our observations, this statement does not always hold true. For example, we showed in Figure 6 that compared to static features, temporal features produced lower reconstruction loss during representation learning, but in Figure 8 we did not observe significant improvement using one-size-fits-all LSTM model to predict mood, health or stress. Similarly, when Jaques et al. [41] used autoencoders to fill in missing values of multimodal features, they found that although the autoencoder loss had been reduced, the prediction rate using the imputed features for binary mood classification was not improved. The authors hypothesized that this was because their data were already rather clean (only 30% training samples contained missing values), and thus their imputation did not show observable effects to prediction. Nevertheless, with personalized prediction models, we demonstrated a definite increase of performance in all three labels where the highest MAE reduction occurred to health by a factor of 13.5%. We also observed that personalizing temporal auto features caused a reduction in the reconstruction loss yet did not improve health and stress prediction.

Similar to Jaques's hypothesis, we hypothesize that this is because our temporal representation model has already captured useful information that can be utilized in the personalization prediction, thus introducing more personalized information to features did not cause an equivalent improvement in prediction results. Therefore, it is likely that the individual difference in subjectively reported scores was dominant over inter-personal differences in passively collected physiological and activity data, thus making personalization the key to transmitting the advantage from features to prediction. More convincingly, the importance of personalization has been proven by the evident gap in performance between one-size-fits-all and MTL prediction approaches.

Numerous studies addressed the same assertion about personalization on a variety of topics including accelerometer-based gesture recognition [54], emotion recognition based on physiological signals [112], etc. Our proposed LC-LSTM-DAE is an unsupervised learning framework that does not require labels to learn low-dimension features. Reducing dimensionality can enhance model generalizability, thus lowering the barrier on the minimal requirement of training set size. We also showed that LC-LSTM-DAE can generalize with smaller partial training sets as well as the full set (Figure 5).

To enable the prediction of unseen participants, we relaxed the personalization strategy using personal profile clustering. It resulted in comparable performance (Mood: 14.5±0.4, health: 14.4±0.4, stress: 15.7±0.5) with the strict individual-as-task personalization (Mood: 14.1±0.2 , health: 12.4±0.3 , stress: 15.0±0.3), in comparison to the one-size-fits-all (Mood: 18.1±0.3, health: 19.3±0.8, stress: 19.9±0.5) scheme. Compared to prior work [42] predicting mood, health, and stress using multimodal features with mean MAE of, respectively, 13.0, 14.1 and 12.9, our reported precision was slightly lower, which was well expected because [42] used features from much more comprehensive modalities including sensors, phone usage, calls/sms, location, weather, and survey, whereas we only used sensor data.

Although the generalized model did not perform as well as personalized models, there were some tactics that could improve the quality of generalized wellbeing learning from auto features. For example, time-step batch normalization, according to Cooijmans et al. [19], improved their results of using several recurrent architectures on a text prediction dataset. Additionally, Laurent et al. [49] showed that, for a speech recognition task, recurrent batch normalization could speed up convergence but not improve generalization performance. In our case, applying time-step batch normalization did not significantly reduce average MAE, but it refined the distribution of predicted scores and their correlations with true scores.

In terms of personal profile clustering, although the overall tendency of Silhouette score and prediction performance matched with each other, higher Silhouette scores (or good clustering of personality and gender) may not always guarantee performance improvement in predicting wellbeing. This could intuitively explain why we observed some fluctuation in both user-dependent and, especially, user-independent cases.

6.2 Interpretation: Understanding Auto Features And Models

Based on observations that the learned features significantly improved wellbeing prediction performance, we have good reason to believe that our proposed representation learning model has successfully learned to extract efficient features from raw sensory data. Hence, it raises a natural question – *what* are those features? To answer this question, we provide an in-depth discussion on the interpretation analysis as follows.

First, we revisit Figure 12 to address the importance of captured information where we count for occurrences of crafted features that provided top-5 correlation coefficients for each auto feature. The more frequent a crafted feature appears strongly correlated with one or more auto features, it would be recognized as being paid special attention by the deep representation model, thus more important than not-so-frequently occurring features. Then and more importantly, by identifying the most important hand-crafted features, we can get a view of what particular kinds of things that the representation model might be looking for.

Our insights of important features were consistent with prior studies that also paid attention to recognizing the contribution of features to wellbeing prediction. Sano et al. [82] selected features to predict self-perceived stress levels via 10-fold cross validation. Specifically, mean, maximum, and median amplitude of SC as well as minimum ST from late morning to evening were the most selected physiological features. We also found that mean SC, median SC, and minimum ST were among the top-5 most frequent strong-correlation providers, but our focal time was before midnight (17H+). Another study [40] focused on predicting happiness from multimodal data ranked the importance of features by information gain, where the authors found that mean, AUC, and S.D. of SC during sleep time could be good indicators of students' happiness. Since our representation model learned deep features in an unsupervised manner whereas the aforementioned [82] and [40] specified the target while selecting features, we may observe discrepancy in those results such as the focal time. For example, we also found that information related with the number of SC peaks was well incorporated by our representation model, whereas other studies did not report it as an important predictor for any label.

In the meantime, we exploited attention mechanisms to help understand the temporal saliency of auto features. We inserted an attention layer into the generalized LSTM wellbeing prediction model, and it could be clearly observed that attention was paid more to days that were coming closer to the wellbeing report time. The finding was intuitive, and it also aligned well with previous studies such as [106]. This indicates that people's current state of wellbeing is likely to be affected by things that occurred several days ago, but the impact would gradually fade away with an accelerated rate.

6.3 Computation, Privacy, And Ethics In Ubiquitous Health Systems

In this study, the mapping from features to predictions could be linearly characterized in a personalized manner, which was computational friendly. Indeed, the representation learning part was time-consuming and energy-heavy. In fact, on a Linux machine with Intel I7-9700k CPU and NVIDIA RTX 2080 Ti GPU, it took more than four hours to train an LC-LSTM-DAE from scratch and over two hours to finetune it to individuals. Nonetheless, the good news was that after the representation model finished training with raw data, it only took 2.3 milliseconds to extract daily features. Moreover, the representation model can be compressed to cost less memory and fewer operations. This is possible because deep learning models are usually sparse and redundant [55]. For example, See et al. [83] proposed magnitude-based weight pruning, reduced the LSTM translation model's size by 90%, yet still kept the performance untouched with re-training. Whereas, Wang et al. [100] pointed out that pruning nodes or connections was limited to undetermined changes in computational paths, and so they proposed efficient implementations of LSTM using structured compression manners that could achieve up to 18.8X and 33.5X gains in speed and energy efficiency. Using similar techniques, Zhang et al. [110] successfully ported a hybrid CNN-LSTM model, which usually ran on multiple GPUs, on an FPGA to recognize video content in real-time. To sum up, existing technologies support that we could compress and distribute our pre-trained representation and prediction models on wearable devices such as a smartwatch. As the sensor data stream in, computation and inference can take place locally in real-time, avoiding data exchange and any risk that it may bring about.

Therefore, making the algorithms more energy efficient not only saves time and battery but also acts as a good way to protect user privacy in mobile health apps [26]. User privacy has been an everlasting concern for any health monitoring application. When it comes to mental health with ubiquitous sensors, privacy and security are even more sensitive due to disagreements among researchers and lack of guidelines [61]. Especially with the rapid growth of interdisciplinary collaboration to leverage deep learning in affective computing, human subject data are usually shared on cloud among multiple parties, bringing up challenges for regulating sensitive data acquisition, management, and usage [115].

Throughout our study, we endeavored to secure participants' privacy from the following aspects. First, this study was conducted strictly under participants' consent. They were free to quit the study at any time. In fact,

several students chose to drop out, with one particularly stating the reason as being concerned about data privacy [78]. Second, all data have been de-identified and stored on our lab server, under information protection measures and policies of the university. Although the LC-LSTM-DAE utilized raw sensitive data and ran on a server, the data were anonymous and encrypted. Above all else, the most privacy-concerning stage, namely the inference of one's wellbeing by linear MTL, can be ported on a smartwatch without effort. Edge computing keeps sensitive data local, thus it can minimize the risk of being exposed to cyber attacks and misuse [74]. Other to-be-considered options to preserve privacy on distributed systems include decentralized computing [23], differential privacy [22], federated computing [89], etc.

6.4 Limitations And Future Work

Our wellbeing prediction based on physiological and behavioral sensors has some limitations. First, how to handle or impute missing data remains an important challenge. We used very simple imputer which merely filled the missing points with mean. With intelligent imputation methods such as Generative Adversarial Nets [104], we might be able to learn more informative features. Although in this paper, we extracted features from three channels of sensor data separately, we will consider combining the multimodal sensor data at early stage and extracting hybrid features in our future work.

Second, some data were not fully utilized due to the absence of self-reported labels. This has been a very common issue in big data and deep learning – the acquisition of ground truth can be labor-heavy, expensive, and sometimes inaccurate. In this study, we simply abandoned unlabeled data, but in the future, we might consider increasing data utilization by semi-supervised learning approaches.

Third, the personalized learning of features needs further investigation. We proved it useful for majority users in terms of learning more robust features, but the bonus was not adequately prompted to prediction. Analysis of the difference between personalized and non-personalized features are in our next steps.

Additionally, one could argue that wearable sensor data might not contain every piece of useful information that relates with future wellbeing. For example, social interactions have been proved to be impactful on mental health [12], but a physiology or motion sensor by itself apparently cannot capture that information as sufficiently as smartphone logs of calls, messages and app usage. Also, if we want to design interventions, the features need to be related with understandable and modifiable behaviors. Unfortunately, most of the time, physiology is not consciously controllable. As a complementary choice, mobile phones can also provide insights into features related with modifiable behaviors such as phone usage, social connections, mobility patterns, and others. With adequate knowledge of causal effects from those features to wellbeing states, accordingly designing recommendations or interventions would be possible. However, we did not include either phone data or causality studies in this paper. Future work can complement sensor data with phone data to learn more comprehensive and powerful features that can be used to train better models of predicting and understanding health or other labels of concern.

Aware of the context-dependent nature of the reported results and our data that came from college students in a NE university, we do not claim reproducible results using the same procedure and principles with another dataset. However, we plan to investigate transfer learning techniques to produce consistent results for other populations, such as non student populations, patients with mental disorders, and other communities in need.

7 CONCLUSION

In this study, we investigated the possibility of automatically learning efficient features from high-resolution time series data passively collected by physiological and behavioral wrist-worn sensors. Our aim was to develop human wellbeing forecast technology that can support personalized long-term health-monitoring and early-warning systems. We collected the dataset (skin conductance, skin temperature, and accelerometer data) and demonstrated that the recurrent LC-LSTM-DAE network can extract better features than feature crafting or static autoencoders.

The results obtained in experiments show that temporal features can be reconstructed to the original raw data dimension with lower reconstruction loss than static features; multi-task models trained on temporal features can predict mood, health, and stress scores with much smaller errors than on crafted features. Another contribution of this paper is the attempt of interpreting deep features and prediction behaviors using correlation analysis, weight visualization, and attention mechanisms. We also provided evidence on the important role that personalization played in predicting subjective targets. We additionally investigated a relaxed personalization strategy that could adapt to unseen users based on some knowledge of their personal profile. The results suggest that although unseen users negatively affect wellbeing prediction, generalization can still be achieved without a critical loss in performance. Finally, we discussed computation, privacy, and ethics. Further studies are recommended to realize the promising potentials of implementing this framework as a real-time ubiquitous system.

ACKNOWLEDGMENTS

This work was supported by NSF (#1840167), NIH (R01GM10518), Samsung Electronics, and NEC Corporation. The study was approved by MIT's COUHES (#1209005240) and Rice University (IRB-FY2018-451). We thank SNAPSHOT study participants and collaborators.

REFERENCES

[1] Hervé Abdi and Lynne J Williams. 2010. Tukey's honestly significant difference (HSD) test. *Encyclopedia of Research Design. Thousand Oaks, CA: Sage* (2010), 1–5.

[2] Mohammed Al-Sarem and Abdel-Hamid Emara. 2019. The effect of training set size in authorship attribution: application on short arabic texts. *International Journal of Electrical and Computer Engineering* 9, 1 (2019), 652.

[3] Salma Alhagry, Aly Aly Fahmy, and Reda A El-Khoribi. 2017. Emotion recognition based on EEG using LSTM recurrent neural network. *Emotion* 8, 10 (2017), 355–358.

[4] Bandar Almaslukh, Jalal AlMuhtadi, and Abdelmonim Artoli. 2017. An effective deep autoencoder approach for online smartphone-based human activity recognition. *Int. J. Comput. Sci. Netw. Secur* 17, 4 (2017), 160–165.

[5] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Vaino Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu. 2016. Deep Speech 2: End-to-end Speech Recognition in English and Mandarin. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16)*. JMLR.org, 173–182. http://dl.acm.org/citation.cfm?id=3045390.3045410

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[7] Mirza Mansoor Baig and Hamid Gholamhosseini. 2013. Smart health monitoring systems: an overview of design and modeling. *Journal of medical systems* 37, 2 (2013), 9898.

[8] Pierre Baldi. 2012. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*. 37–49.

[9] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, Alex, and Pentland. 2014. Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits. *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia* (10 2014). https://doi.org/10.1145/2647868.2654933

[10] Wolfram Boucsein. 2012. *Electrodermal activity*. Springer Science & Business Media.

[11] Rich Caruana. 1993. Multitask Learning: A Knowledge-Based Source of Inductive Bias. In *ICML*.

[12] Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. 2019. Who is the" Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–32.

[13] Tanaya Chaudhuri, Deqing Zhai, Yeng Chai Soh, Hua Li, and Lihua Xie. 2018. Thermal comfort prediction using normalized skin temperature in a uniform built environment. *Energy and Buildings* 159 (2018), 426–440.

[14] Weixuan Chen, Natasha Jaques, Sara Taylor, Akane Sano, Szymon Fedor, and Rosalind W Picard. 2015. Wavelet-based motion artifact removal for electrodermal activity. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 6223–6226.
[15] Davide Chicco, Peter Sadowski, and Pierre Baldi. 2014. Deep autoencoder neural networks for gene ontology annotation predictions. In *Proceedings of the 5th ACM conference on bioinformatics, computational biology, and health informatics*. 533–540.
[16] Dami Choi, Alexandre Passos, Christopher J Shallue, and George E Dahl. 2019. Faster neural network training with data echoing. *arXiv preprint arXiv:1907.05550* (2019).
[17] Russell M Church. 1962. The effects of competition on reaction time and palmar skin conductance. *The Journal of Abnormal and Social Psychology* 65, 1 (1962), 32.
[18] Ana Ciocarlan, Judith Masthoff, and Nir Oren. 2018. Kindness is contagious: Study into exploring engagement and adapting persuasive games for wellbeing. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. ACM, 311–319.
[19] Tim Cooijmans, Nicolas Ballas, César Laurent, Çağlar Gülçehre, and Aaron Courville. 2016. Recurrent batch normalization. *arXiv preprint arXiv:1603.09025* (2016).
[20] Paul De Bra. 2017. Challenges in user modeling and personalization. *IEEE Intelligent Systems* 32, 5 (2017), 76–80.
[21] Halbert L Dunn. 1959. High-level wellness for man and society. *American journal of public health and the nations health* 49, 6 (1959), 786–792.
[22] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
[23] Zipei Fan, Xuan Song, Renhe Jiang, Quanjun Chen, and Ryosuke Shibasaki. 2019. Decentralized Attention-based Personalized Human Mobility Prediction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–26.
[24] Enrique Garcia-Ceja, Venet Osmani, and Oscar Mayora. 2015. Automatic stress detection in working environments from smartphones' accelerometer data: a first step. *IEEE journal of biomedical and health informatics* 20, 4 (2015), 1053–1060.
[25] Fabio Gasparetti. 2017. Personalization and context-awareness in social local search: State-of-the-art and future research challenges. *Pervasive and Mobile Computing* 38 (2017), 446–473.
[26] Munkhjargal Gochoo, Tan-Hsu Tan, Shih-Chia Huang, Tsedevdorj Batjargal, Jun-Wei Hsieh, Fady S Alnajjar, and Yung-Fu Chen. 2019. Novel IoT-Based Privacy-Preserving Yoga Posture Recognition System Using Low-Resolution Infrared Sensors and Deep Learning. *IEEE Internet of Things Journal* (2019).
[27] Joseph F Hair, William C Black, Barry J Babin, Rolph E Anderson, Ronald L Tatham, et al. 1998. *Multivariate data analysis*. Vol. 5. Prentice hall Upper Saddle River, NJ.
[28] Nils Yannick Hammerla, James Fisher, Peter Andras, Lynn Rochester, Richard Walker, and Thomas Plötz. 2015. PD disease state assessment in naturalistic environments using deep learning. In *Twenty-Ninth AAAI conference on artificial intelligence*.
[29] Harish Haresamudram, David V Anderson, and Thomas Plötz. 2019. On the role of features in human activity recognition. In *Proceedings of the 23rd International Symposium on Wearable Computers*. 78–88.
[30] Jennifer A Healey and Rosalind W Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems* 6, 2 (2005), 156–166.
[31] Robert Hecht-Nielsen. 1992. Theory of the backpropagation neural network. In *Neural networks for perception*. Elsevier, 65–93.
[32] Fabio Hernández, Luis F Suárez, Javier Villamizar, and Miguel Altuve. 2019. Human Activity Recognition on Smartphones Using a Bidirectional LSTM Network. In *2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA)*. IEEE, 1–5.
[33] Javier Hernandez, Rob R Morris, and Rosalind W Picard. 2011. Call center stress recognition with person-specific models. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 125–134.
[34] HM Sajjad Hossain and Nirmalya Roy. 2019. Active Deep Learning for Activity Recognition with Context Aware Annotator Selection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1862–1870.
[35] Karen Hovsepian, Mustafa Al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 493–504.
[36] Jason Hsu. 1996. *Multiple comparisons: theory and methods*. Chapman and Hall/CRC.
[37] Zhexue Huang. 1997. Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*. Singapore, 21–34.
[38] Gudmund R Iversen, Albert R Wildt, Helmut Norpoth, and Helmut P Norpoth. 1987. *Analysis of variance*. Number 1. Sage.
[39] Natasha Jaques, Ognjen (Oggi) Rudovic, Sara Taylor, Akane Sano, and Rosalind Picard. 2017. Predicting Tomorrow's Mood, Health, and Stress Level using Personalized Multitask Learning and Domain Adaptation. In *Proceedings of IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing (Proceedings of Machine Learning Research)*, Neil Lawrence and Mark Reid (Eds.), Vol. 66. PMLR, 17–33.
[40] Natasha Jaques, Sara Taylor, Asaph Azaria, Asma Ghandeharioun, Akane Sano, and Rosalind Picard. 2015. Predicting students' happiness from physiology, phone, mobility, and behavioral data. In *2015 International Conference on Affective Computing and Intelligent*

Interaction (ACII). IEEE, 222–228.

[41] Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. 2017. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 202–208.

[42] Natasha Jaques, Sara Taylor, Akane Sano, Rosalind Picard, et al. 2017. Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation. In *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*. 17–33.

[43] Oliver P. John and Sanjay Srivastava. 1999. The Big-five Trait Taxonomy: History, Measurement, and Theoretical Perspectives.

[44] Eiman Kanjo, Eman MG Younis, and Chee Siang Ang. 2019. Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information Fusion* 49 (2019), 46–56.

[45] Ronald C Kessler, Patricia A Berglund, Martha L Bruce, J Randy Koch, Eugene M Laska, Philip J Leaf, Ronald W Manderscheid, Robert A Rosenheck, Ellen E Walters, and Philip S Wang. 2001. The prevalence and correlates of untreated serious mental illness. *Health services research* 36, 6 Pt 1 (2001), 987.

[46] Jinkyu Kim and John Canny. 2017. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision*. 2942–2950.

[47] Mark A Kramer. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal* 37, 2 (1991), 233–243.

[48] K Krauchi and ANNA Wirz-Justice. 1994. Circadian rhythm of heat production, heart rate, and skin and core temperature under unmasking conditions in men. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 267, 3 (1994), R819–R829.

[49] César Laurent, Gabriel Pereyra, Philémon Brakel, Ying Zhang, and Yoshua Bengio. 2016. Batch normalized recurrent neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2657–2661.

[50] Boning Li, Han Yu, and Akane Sano. 2019. Toward End-to-end Prediction of Future Wellbeing using Deep Sensor Representation Learning. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 253–257.

[51] Boning Li, Han Yu, and Akane Sano. 2019. Toward End-to-end Prediction of Future Wellbeing using Deep Sensor Representation Learning. *Machine Learning for the Diagnosis and Treatment of Affective Disorders, ACII Workshop* (2019), London UK.

[52] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. 2013. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. 389–402.

[53] A. Liu, Y. Su, W. Nie, and M. Kankanhalli. 2017. Hierarchical Clustering Multi-Task Learning for Joint Human Action Grouping and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 1 (Jan 2017), 102–114. https://doi.org/10.1109/TPAMI.2016.2537337

[54] Jiayang Liu, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. 2009. uWave: Accelerometer-based personalized gesture recognition and its applications. *Pervasive and Mobile Computing* 5, 6 (2009), 657–675.

[55] Sicong Liu, Yingyan Lin, Zimu Zhou, Kaiming Nan, Hui Liu, and Junzhao Du. 2018. On-demand deep model compression for mobile devices: A usage-driven model selection framework. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 389–400.

[56] Yongshuai Liu, Jiyu Chen, and Hao Chen. 2018. Less is more: Culling the training set to improve robustness of deep neural networks. In *International Conference on Decision and Game Theory for Security*. Springer, 102–114.

[57] Jin Lu, Chao Shang, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jayesh Kamath, Athanasios Bamis, Alexander Russell, Bing Wang, and Jinbo Bi. 2018. Joint Modeling of Heterogeneous Sensing Data for Depression Assessment via Multi-task Learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 21 (March 2018), 21 pages. https://doi.org/10.1145/3191753

[58] Abhinav Mehrotra and Mirco Musolesi. 2018. Using autoencoders to automatically extract mobility features for predicting depressive states. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 127.

[59] Kathleen Ries Merikangas, Joel Swendsen, Ian B Hickie, Lihong Cui, Haochang Shou, Alison K Merikangas, Jihui Zhang, Femke Lamers, Ciprian Crainiceanu, Nora D Volkow, et al. 2019. Real-time mobile monitoring of the dynamic associations among motor activity, energy, mood, and sleep in adults with bipolar disorder. *JAMA psychiatry* 76, 2 (2019), 190–198.

[60] Varun Mishra. 2019. From sensing to intervention for mental and behavioral health. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 388–392.

[61] David C. Mohr, Mi Zhang, and Stephen M. Schueller. 2017. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual Review of Clinical Psychology* 13, 1 (2017), 23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949 arXiv:https://doi.org/10.1146/annurev-clinpsy-032816-044949 PMID: 28375728.

[62] Jill K Morris, Robyn A Honea, Eric D Vidoni, Russell H Swerdlow, and Jeffrey M Burns. 2014. Is Alzheimer's disease a systemic disease? *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1842, 9 (2014), 1340–1349.

[63] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K D'Mello, Munmun De Choudhury, Gregory D Abowd, and Thomas Plötz. 2019. Prediction of Mood Instability with Passive Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 75.

[64] Chung Wai Mark Ng, Choon How How, and Yin Ping Ng. 2016. Major depression in primary care: making the diagnosis. *Singapore medical journal* 57, 11 (2016), 591.

[65] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. 2015. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 685–694.

[66] Francisco Ordóñez and Daniel Roggen. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16, 1 (2016), 115.

[67] K Palanisamy, M Murugappan, and S Yaacob. 2013. Multiple physiological signal-based human stress identification using non-linear classifiers. *Elektronika ir elektrotechnika* 19, 7 (2013), 80–85.

[68] Liangying Peng, Ling Chen, Zhenan Ye, and Yi Zhang. 2018. Aroma: A deep multi-task learning based simple and complex human activity recognition method using wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–16.

[69] Thomas Plötz, Nils Y Hammerla, and Patrick L Olivier. 2011. Feature learning for activity recognition in ubiquitous computing. In *Twenty-second international joint conference on artificial intelligence*.

[70] Kevin L. Priddy and Paul E. Keller. 2005. *Artificial Neural Networks: An Introduction (SPIE Tutorial Texts in Optical Engineering, Vol. TT68)*. SPIE- International Society for Optical Engineering.

[71] Yeping Lina Qiu, Hong Zheng, and Olivier Gevaert. 2018. A deep learning framework for imputing missing values in genomic data. *bioRxiv* (2018), 406066.

[72] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*. 6076–6085.

[73] Victoria L Richmond, Sarah Davey, Katy Griggs, and George Havenith. 2015. Prediction of core body temperature from multiple variables. *Annals of occupational hygiene* 59, 9 (2015), 1168–1178.

[74] Rodrigo Roman, Javier Lopez, and Masahiro Mambo. 2018. Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges. *Future Generation Computer Systems* 78 (2018), 680–698.

[75] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

[76] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research* 17, 7 (2015), e175.

[77] Mayu Sakurada and Takehisa Yairi. 2014. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*. ACM, 4.

[78] Akane Sano. 2016. *Measuring college students' sleep, stress, mental health and wellbeing with wearable sensors and mobile phones*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[79] Akane Sano, Weixuan Chen, Daniel Lopez-Martinez, Sara Taylor, and Rosalind W Picard. 2018. Multimodal Ambulatory Sleep Detection Using LSTM Recurrent Neural Networks. *IEEE journal of biomedical and health informatics* (2018).

[80] Akane Sano and Rosalind W Picard. 2013. Stress recognition using wearable sensors and mobile phones. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 671–676.

[81] Akane Sano, Rosalind W Picard, and Robert Stickgold. 2014. Quantitative analysis of wrist electrodermal activity during sleep. *International Journal of Psychophysiology* 94, 3 (2014), 382–389.

[82] Akane Sano, Sara Taylor, Andrew W McHill, Andrew JK Phillips, Laura K Barger, Elizabeth Klerman, and Rosalind Picard. 2018. Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: Observational study. *Journal of medical Internet research* 20, 6 (2018), e210.

[83] Abigail See, Minh-Thang Luong, and Christopher D Manning. 2016. Compression of neural machine translation models via pruning. *arXiv preprint arXiv:1606.09274* (2016).

[84] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 806–813.

[85] Dimitris Spathis, Sandra Servia-Rodriguez, Katayoun Farrahi, Cecilia Mascolo, and Jason Rentfrow. 2019. Sequence Multi-task Learning to Forecast Mental Wellbeing from Sparse Self-reported Data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. ACM, New York, NY, USA, 2886–2894. https://doi.org/10.1145/3292500.3330730

[86] Julian Steil, Philipp Müller, Yusuke Sugano, and Andreas Bulling. 2018. Forecasting user attention during everyday mobile interactions using device-integrated and wearable sensors. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–13.

[87] Sara Ann Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2017. Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Transactions on Affective Computing* (2017).

[88] Catherine Tong, Matthew Craner, Matthieu Vegreville, and Nicholas D Lane. 2019. Tracking Fatigue and Health State in Multiple Sclerosis Patients Using Connnected Wellness Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–19.

[89] Stacey Truex, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, Rui Zhang, and Yi Zhou. 2019. A hybrid approach to privacy-preserving federated learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*. 1–11.

[90] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing* 11, 8 (2017), 1301–1309.

[91] Terumi Umematsu, Akane Sano, and Rosalind W Picard. 2019. Daytime Data and LSTM can Forecast Tomorrow's Stress, Health, and Happiness. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2186–2190.

[92] Terumi Umematsu, Akane Sano, Sara Taylor, and Rosalind W. Picard. 2019. Improving Students' Daily Life Stress Forecasting using LSTM Neural Networks. *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)* (2019), 1–4.

[93] Terumi Umematsu, Akane Sano, Sara Taylor, and Rosalind W Picard. 2019. Improving Students' Daily Life Stress Forecasting using LSTM Neural Networks. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 1–4.

[94] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. 2009. Dimensionality reduction: a comparative. *J Mach Learn Res* 10, 66-71 (2009), 13.

[95] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. 2014. scikit-image: image processing in Python. *PeerJ* 2 (2014), e453.

[96] Peilu Wang, Yao Qian, Frank K Soong, Lei He, and Hai Zhao. 2015. A unified tagging solution: Bidirectional lstm recurrent neural network with word embedding. *arXiv preprint arXiv:1511.00215* (2015).

[97] Rui Wang, Min SH Aung, Saeed Abdullah, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill, Emily A Scherer, et al. 2016. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 886–897.

[98] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. 3–14.

[99] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. 2018. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–26.

[100] Shuo Wang, Zhe Li, Caiwen Ding, Bo Yuan, Qinru Qiu, Yanzhi Wang, and Yun Liang. 2018. C-LSTM: Enabling efficient LSTM using structured compression techniques on FPGAs. In *Proceedings of the 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. 11–20.

[101] Genta Indra Winata, Onno Pepijn Kampman, and Pascale Fung. 2018. Attention-based lstm for psychological stress detection from spoken language using distant supervision. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6204–6208.

[102] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[103] Takehiro Yamakoshi, K Yamakoshi, S Tanaka, M Nogawa, Sang-Bum Park, Mariko Shibata, Y Sawada, P Rolfe, and Yasuo Hirose. 2008. Feasibility study on driver's stress detection from differential skin temperature measurement. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 1076–1079.

[104] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. 2018. Gain: Missing data imputation using generative adversarial nets. *arXiv preprint arXiv:1806.02920* (2018).

[105] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015).

[106] Han Yu, EB Klerman, Rosalind Picard, and Akane Sano. 2019. Personalized Wellbeing Prediction using Behavioral, Physiological and Weather Data. *IEEE-EMBS Biomedical and Health Informatics 2019* (2019).

[107] Ming Zeng, Haoxiang Gao, Tong Yu, Ole J Mengshoel, Helge Langseth, Ian Lane, and Xiaobing Liu. 2018. Understanding and improving recurrent networks for human activity recognition by continuous attention. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*. 56–63.

[108] Alexandros Zenonos, Aftab Khan, Georgios Kalogridis, Stefanos Vatsikas, Tim Lewis, and Mahesh Sooriyabandara. 2016. HealthyOffice: Mood recognition at work using smartphones and wearable sensors. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. IEEE, 1–6.

[109] Fusang Zhang, Kai Niu, Jie Xiong, Beihong Jin, Tao Gu, Yuhang Jiang, and Daqing Zhang. 2019. Towards a Diffraction-based Sensing Approach on Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 33 (March 2019), 25 pages. https://doi.org/10.1145/3314420

[110] Xiaofan Zhang, Xinheng Liu, Anand Ramachandran, Chuanhao Zhuge, Shibin Tang, Peng Ouyang, Zuofu Cheng, Kyle Rupnow, and Deming Chen. 2017. High-performance video content recognition with long-term recurrent convolutional network for FPGA. In *2017 27th International Conference on Field Programmable Logic and Applications (FPL)*. IEEE, 1–4.

[111] Xiao-Ping Zhang and Mita D Desai. 1998. Adaptive denoising based on SURE risk. *IEEE signal processing letters* 5, 10 (1998), 265–267.

[112] Sicheng Zhao, Amir Gholaminejad, Guiguang Ding, Yue Gao, Jungong Han, and Kurt Keutzer. 2019. Personalized Emotion Recognition by Personality-Aware High-Order Learning of Physiological Signals. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1s, Article 14 (Jan. 2019), 18 pages. https://doi.org/10.1145/3233184

[113] Wenliang Zhong and James Kwok. 2012. Convex multitask learning with flexible task clusters. *arXiv preprint arXiv:1206.4601* (2012).

[114] Jiayu Zhou, Jianhui Chen, and Jieping Ye. 2011. Malsar: Multi-task learning via structural regularization. *Arizona State University* 21 (2011).

[115] Zhenyu Zhou, Haijun Liao, Bo Gu, Kazi Mohammed Saidul Huq, Shahid Mumtaz, and Jonathan Rodriguez. 2018. Robust mobile crowd sensing: When deep learning meets edge computing. *IEEE Network* 32, 4 (2018), 54–60.