RICE UNIVERSITY

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

By

Boning Li

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

Master of Science

APPROVED, THESIS COMMITTEE

Akane Sano

Akane Sano
Assistant Professor of Electrical and Computer Engineering and Computer Science

Ashutosh Sabharwal
Professor of Electrical and Computer Engineering

Ashok Veeraraghavan
Associate Professor of Electrical and Computer Engineering and Computer Science

HOUSTON, TEXAS
August 2020
ABSTRACT

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

by

Boning Li

High-resolution wearable sensor data contain physiological and behavioral information that can be utilized to predict and eventually improve human health and well-being. We propose a semi-supervised deep neural network framework to automatically learn features from passively collected multi-modal sensor data. This process can be personalized by finetuning the general features with participant-specific data. Then, using the learned features, we performed personalized prediction of subjective well-being scores with high precision. We also provide visual explanation and statistical interpretation of the automatically learned features and the prediction models. In this study, we explored multiple implementations of our framework including locally connected linear network, convolutional neural network, recurrent neural network, and visual attention network. 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 scores include self-reported mood, health and stress ranged from 0 to 100. Compared to the prediction performance based on hand-crafted features, the proposed framework achieved higher precision with a smaller number of features. Our results show promising potentials 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.
Acknowledgements

I would like to thank my thesis advisor Professor Akane Sano of the Electrical and Computer Engineering Department at Rice University. Professor Sano was always available and responsive to my research questions. Her continued guidance and insightful suggestions were fundamental to the formulating of this thesis topic, directions, and methodology.

I wish to thank my thesis committee members, Professor Ashutosh Sabharwal and Professor Ashok Veeraraghavan for the time, support, guidance, and kindness that they have generously offered me during the preparation of this thesis.

I would also like to acknowledge our sponsors NSF (#1840167), NIH (R01GM10518), Samsung Electronics, and NEC Corporation for supporting this research. Same acknowledgement goes to SNAPSHOT Study participants and collaborators. The dataset to which they contributed was essential to this study.

Finally, I must express my immense gratitude to my parents and to my friends who offered unfailing encouragement and never-ending support to me throughout the years. I would have never reached this milestone without them during this particularly difficult period. Thank you.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>List of Illustrations</td>
<td>ix</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xiii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2 Related Work</td>
<td>5</td>
</tr>
<tr>
<td>2.1 Wellbeing Prediction through Passive Sensing</td>
<td>5</td>
</tr>
<tr>
<td>2.2 Deep Learning: Prediction, Representation, and Interpretation</td>
<td>6</td>
</tr>
<tr>
<td>2.3 Personalization in Feature Extraction and Prediction</td>
<td>9</td>
</tr>
<tr>
<td>3 Methods</td>
<td>11</td>
</tr>
<tr>
<td>3.1 Data Collection</td>
<td>11</td>
</tr>
<tr>
<td>3.2 Feature Extraction</td>
<td>14</td>
</tr>
<tr>
<td>3.2.1 Baseline Feature Crafting</td>
<td>14</td>
</tr>
<tr>
<td>3.2.2 Deep Feature Learning</td>
<td>14</td>
</tr>
<tr>
<td>3.2.3 Modality Fusion</td>
<td>20</td>
</tr>
<tr>
<td>3.2.4 Semi-supervised Learning</td>
<td>21</td>
</tr>
<tr>
<td>3.3 Wellbeing Prediction</td>
<td>22</td>
</tr>
<tr>
<td>3.3.1 Generalized Model – One-size-fits-all LSTM with Recurrent Batch Normalization</td>
<td>22</td>
</tr>
<tr>
<td>3.3.2 Personalized Model – Multi-task Linear Regularized Models</td>
<td>23</td>
</tr>
<tr>
<td>3.3.3 Evaluation Metrics</td>
<td>24</td>
</tr>
<tr>
<td>3.4 Interpretation</td>
<td>26</td>
</tr>
</tbody>
</table>
5.1.5 Auto-learned vs Hand-crafted Features: Prediction Performance Comparison 43

5.1.6 Generalized Multi-day LSTM Wellbeing Models 43

5.1.7 Clustering Based on Profile 44

5.1.8 User Dependent vs Independent Policies 45

5.1.9 Interpretation of Features and Predictions 47

5.2 Late vs Early Modality Fusion 51

5.2.1 Representation Learning 51

5.2.2 Wellbeing Prediction 52

5.2.3 Critical Feature Analysis 53

5.3 Semi-Supervised Framework Based Results 54

5.3.1 Representation Learning 55

5.3.2 Wellbeing Prediction 56

5.3.3 Statistical Interpretation 59

5.4 Semi-Supervised Convolutional Visual Attention Map 63

6 Discussion 69

6.1 From Features to Wellbeing Labels 69

6.2 When to Fuse Which Modalities 72

6.3 The Role of Supervision in Representation Learning 73

6.4 Understanding Auto Features and Deep Models 75

6.5 Attention is Not All We Need 76

6.6 Computation, Privacy, and Ethics in Ubiquitous Health Systems 78

6.7 Limitations and Future Work 80

7 Conclusion 82

A Full List of Crafted Baseline Features 84
Illustrations

3.1 Self-reported mood (sad-happy), health (sick-healthy), and stress (stressed-calm) scores. .............................................. 13

3.2 Schematics of the representation learning framework for automatic feature extraction from raw sensor data. In this illustration, we selected locally connected MLP as the static encoder, and LSTM as the temporal encoder at the bottleneck. In this study, we also explored 1D CNN as the static encoder, and (recurrent) attention layers as the temporal encoder. The inference module generating wellbeing scores from the learned features served as supervision under the semi-supervised settings, and it should be disconnected under the unsupervised settings. .............................................. 15

3.3 Illustration of modality fusion schemes. ................................. 21

3.4 Illustration of attention mechanism on top of a many-to-one LSTM network, where $h_t$ denotes the hidden state at time-step $t$; $\alpha$ is the attention weight; $u$ and $v$ are the importance vector and the context vector. .............................................. 27

5.1 Reconstruction loss on validation set with different input data resolution for each channel. .............................................. 41

5.2 Reconstruction loss vs training-set size. .................................. 42

5.3 Reconstruction loss vs autoencoder architecture. ....................... 42
5.4 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 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.

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

5.7 Silhouette score vs number of user groups.

5.8 Wellbeing prediction MAE vs number of user groups.

5.9 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 × 10⁻⁵). The auto SC, ST and AC features were independently learned from the corresponding modalities and were not personalized.

5.10 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.

5.11 Comparison of the unsupervised representation learning loss on the timing of fusing three modalities (skin conductance, skin temperature, and acceleration). The losses were computed on the validation set during a iterative mini-batch training session.
5.12 Performance of wellbeing prediction using user-dependent and user-independent personalizing MTL on different fusion schemes and modality combinations. For the fused modalities, the inner bars denote early fusion; the outer bars denote late fusion.

5.13 Semi-supervised learning curves of weighted inference loss, weighted reconstruction loss, and total loss on the validation set. Different y-axes were used to plot reconstruction loss (right vertical axis) and inference loss (left vertical axis) in order for better illustration. The total loss is equal to the sum of the weighted reconstruction and inference losses ($\lambda = 0.5$).

5.14 Personalized health prediction performance could be strongly impacted by the variability in ground-truth health scores. Higher prediction performance was achieved with more stable participants.

5.15 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 $\pm 0.05$ ($p < 1.7 \times 10^{-5}$). The auto SC, ST and AC features were independently learned from the corresponding modalities and were not personalized.

5.16 Semi-supervised learning curves of the convolutional visual attention model based on the SC data and mood, health, and stress scores. The total loss is equal to the sum of weighted reconstruction and inference losses ($\lambda = 0.5$).

5.17 Segment of an upsampling example of attention weights by Gaussian RBF interpolation, with and without max pooling indices.
5.18 Upsampling convolutional visual attention with max-unpooling. (a) displays the original input SC data in the top plot, and the bottom plot is the base attention weights; vertical dash lines separates the 100 base attention weights, and we call the position of each attention weight an attention moment. (b) shows the base attention weights as a line plot and a heatbar. (c) shows the completed upsampling of the plots of (b), resulted in 86,280 upsampled attention moments. The upsampling rate from (b) to (c) was x862.8.

5.19 An instance of the prominence peak detection results on an upsampled attention map.

5.20 Upsampling ST convolutional visual attention with max-unpooling. Prominence peaks are annotated in (c).

5.21 Upsampling AC convolutional visual attention with max-unpooling. Prominence peaks are annotated in (c).

5.22 Ten-second SC data segments centered at high-attention moments.
Tables

3.1 Overview of the dataset .................................................. 12

5.1 Wellbeing prediction mean absolute errors (MAE) using
individual-as-task MTL on different feature sets \(\text{(Mean±S.D.)}\). ........ 44

5.2 Most interpreted auto features for each sensor modality, as well as the
top-5 highest correlated crafted features with associated correlation
coefficients \(p<2.3\times10^{-5}\). .................................................. 48

5.3 Wellbeing prediction performance MAE \(\text{(S.D.)}\) by the automatically
learned features via semi-supervised vs unsupervised representation
learning. An asterisk indicates a statistically significant advantage in
performance \(p<0.05\). .................................................. 57

5.4 Accuracy of tomorrow’s health forecast for a change of self-reported
sick-healthy scores that is at least \(|dy|\) different tomorrow compared
to today, based on generalized and personalized prediction models
with 7-day features. Asterisks denote how significantly that our
predicted transition accuracy outperformed the corresponding
baseline result by McNemar Test \(\ast p<0.05; \ast\ast p<0.01; \ast\ast\ast p<
0.005 ; \ast\ast\ast\ast p<0.001\). .................................................. 59
Chapter 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 [1]. 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 [2]. 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 [3]. If daily health and wellbeing 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 [4]. 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 tasks [5, 6]. 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 [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26].

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 various types of raw data in either unsupervised [27, 28, 29, 30, 31, 32, 33, 34] or (semi-)supervised manner [35, 36, 37, 38, 39]. However, a major drawback lies in the lack of clear explanation of those features and predictions. Although some mechanisms exist to visualize the critical features or steps with slightly more intuitions (e.g., saliency maps and attention mechanism [40, 41]), 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. 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.

Personalization is another understudied factor in detecting self-reported targets. On one hand, individual differences can cause high intra-class variation relative to inter-class variation, thus degrading the performance of a unified model on data from different users [42, 43]. On the other hand, customizing a model for each participant [44] requires a lot of labeled data which are expensive to acquire. Moreover, unless retrained, the model is unable to accommodate unseen users, i.e. who were not originally in the training set. To overcome these shortcomings, clustered multi-task personalization can be utilized with a compromise on prediction accuracy [45, 46, 20].

In this thesis, we propose an interpretable semi-supervised two-stage deep learning framework consisting of 1) a recurrent autoencoder that learns to extract physiological
and behavioral features from raw skin conductance (SC), skin temperature (ST) and acceleration (AC) data, and 2) 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,

i) **Automatic learning of time-series representation from wearable sensors’ raw data.** 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. It can exploit known labels by enabling its semi-supervised mode.

ii) **Semi-supervised convolutional visual attention learning.** We leverage one-dimensional convolutional neural network and attention mechanism to produce attention scores for consecutive timestamps of learned representations. Peaks of attention scores can reveal locations that the model finds important to learn. We also adopt indexed upsampling to increase the resolution of the scores from \(\sim 15\) minutes to \(\sim 1\) second (upsampling rate x862.8).

iii) **Interpretation of deeply learned features and predictions.** We provide interpretation on multiple levels. We leverage attention layers to visualize low-level data 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 linear MTL models from a statistical point of view.

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 visual explanation and statistical interpretation 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. Some chapters in this thesis were adapted to my previous papers [22], [47], and [48].
Chapter 2

Related Work

2.1 Wellbeing Prediction through Passive Sensing

Wellbeing refers to aspects of emotions, mood, mental and physical health \[49\]. 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 \[9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 8, 7, 22, 23, 24, 25, 26\]. For example, Zenonos et al. \[13\] 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. \[14\] 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 targeted to infer various health conditions, such as Schizophrenia symptoms \[15\] and fatigue \[17\]. In \[7\], 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 \[50, 51\]. It has been shown that stress could be assessed by skin conductance \[52\], skin temperature \[53\], and acceleration \[54\]. Sano et al. \[18\] 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, [20] 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 [7], 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. In this study, we 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

An artificial neural network (NN) is based on one or more layers of nodes and connections (i.e. linear transform followed by pointwise nonlinearity), which mimics the learning process of biological neurons or brains [36]. Depending on the architecture and dynamics, the basic concept of NN 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 [56, 57, 58]. Generally speaking, MLP is the “vanilla” form of NN, commonly used on relatively simple datasets; CNN and RNN encompass more specificity, respectively in visual and sequential data [59]. Recently, 1D CNN has been also applied to time series data such as speech and sensors for health related monitoring and prediction [60, 61, 62]. Compared with model-based methods, data-driven deep learning can exploit big data and computational power to learn very complicated mappings, letting alone the parameters of the physical systems. Nonetheless, as Tidriri et al. pointed out in [63], a marriage of the two may result in more favorable performance than either one on its own.

Apart from end-to-end classification or regression, deep learning can also be leveraged to learn complicated features from a large amount of raw input data via autoencoders, in contrast to crafting features with fixed rules. An autoencoder is trained to accurately recover the input data through an encoder-feature-decoder structure, in either unsupervised, supervised, or hybrid manner [64]. A single-layer autoencoder with linear activation is equivalent to principal component analysis (PCA) which is essentially a linear transformation [31]. In most autoencoder practice, however, pointwise non-linear activation functions are inserted to introduce greater expressive power [65]. Also, fully connected layers, convolutional layers, and recurrent layers can be stacked to make it capable of modelling more complicated functions. Consequently, the power comes with a cost such that PCA is faster and computationally cheaper than autoencoderS. Due to the high number of parameters, the good performance of autoencoderS relies on sufficient training data, appropriate regularization, and careful design [66]. Many studies have been carried out to explore the feasibility of distributing raw sensory data in autoencoderS so that wellbeing-related features...
can be automatically learned through back propagation in either unsupervised or supervised manner \cite{30,67}. For example, Mehrotra et al. \cite{32} 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, \cite{33} integrated CNN and long short-term memory (LSTM) networks for human activity recognition based on 500 ms data segmentation from accelerometers and gyroscopes, and \cite{33} 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 proved 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. Different from these existing works where a hard end-to-end prediction model was usually presented, we propose a two-step deep learning framework consisting of a semi-supervised representation model and a supervised prediction model. Our proposed models can leverage high-quality ground truth labels, while accommodating for situations where the labels are noisy or lack.

An often attacked drawback of deep 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 \cite{68,69}. For sequential data, attention mechanisms can determine time-step saliency in RNN models, such as in image captioning \cite{70} and steering angle prediction for self-driving cars \cite{71}. Raghu et al. \cite{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 attention-based 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 [13, 73, 74, 20]. To yield more accurate individual-level stress predictions, Rozet et al. [75] combined generalized and personalized models of physical activities (e.g. Fitbit measured number of steps taken, calories burned, etc.). Unfortunately, personalization is often understudied with regard to the automatically learned features due to the lack of labeled data in research or the lack of prior knowledge of incoming samples in practice [76, 77].

To overcome these defects, clustered multi-task personalization based on personal profile criteria can be an alternative with a compromise on prediction accuracy [45]. Utilizing knowledge of personality types, Ciocarlan et al. [78] 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. [7] 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 always a unified process for all individuals. Motivated by the performance improvement brought by personalizing prediction models, we expect that with adequate data, personalizing the representation model may as well result in higher feature quality. In essence, personalized learning is to utilize knowledge acquired for one dataset to solve another where the data may have a reasonably different distribution. To achieve that, a machine learning technique called transfer learning is commonly used. However, the complete transfer learning diagram of pretraining-retraining-finetuning typically takes many iterations to train and overfits to relatively close datasets. Depending on specific datasets and purposes, researchers can drop certain steps to speed up the training process yet still maintain good performance. For example, implemented personalized wellbeing prediction by sharing low-level layer parameters and directly training separate groups of high-level layers for individual tasks. In this study, on top of a non-personalized 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.
Chapter 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 [80]. 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 [81], and thus it can be an index of stress response [82]. ST can be reflective of many health problems (e.g. fever and heat exhaustion) [83, 84] and biological process (e.g. circadian rhythm and sleep) [85]. AC directly monitors human movement and sleep patterns, thus suitable for measuring energy and physical activities from which mental status could be inferred [11]. 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 [86]. Table 3.1 summarizes 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 3.1.

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

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 [87, 88]. A wavelet filter was adopted to remove artifacts from ST data [89], using Symlets 4 Scaling function and adaptive threshold by Stein’s Unbiased Risk Estimate [90]. For the AC
Figure 3.1: Self-reported mood (sad-happy), health (sick-healthy), and stress (stressed-calm) scores.

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 [91].

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 [7, 20]. More precisely speaking, we divided each day (0-24H) into four non-overlapping time windows, namely 0-3H, 3-10H, 10-17H, and 17-24H. From each time window, 34 unique features were computed, resulting in totally 136 crafted features per day. The full list of 34 unique crafted features can be found in Appendix A. 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 the automatically learned features.

3.2.2 Deep Feature Learning

Overview of the Framework Architecture. We propose the Semi-supervised Locally Connected Recurrent Denoising Autoencoder as the fundamental deep representation learning framework. It is to be trained on daily raw wearable sensor data with or without true wellbeing scores of the next day, in order to learn and extract physiological and behavioral features. In the current section, we set the focus on elaborating the unsupervised components of the representation framework. In Section 3.2.4, we will detail the supervised component.

We composed locally connected layers (i.e. static) and recurrent layers (i.e. tem-
with a denoising autoencoder (DAE) to form this multi-modal representation learning diagram. To learn features that are tuned to a specific task, the framework can easily turn semi-supervised by adding an NN-based inference model mapping the learned features to ground truth labels.

Figure 3.2: Schematics of the representation learning framework for automatic feature extraction from raw sensor data. In this illustration, we selected locally connected MLP as the static encoder, and LSTM as the temporal encoder at the bottleneck. In this study, we also explored 1D CNN as the static encoder, and (recurrent) attention layers as the temporal encoder. The inference module generating wellbeing scores from the learned features served as supervision under the semi-supervised settings, and it should be disconnected under the unsupervised settings.

The framework is illustrated in Figure 3.2. 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 same hierarchical local structures as the static encoder. The static encoder can take the form of various architectures, such as MLP, CNN, etc. Outputs of eight time frames were later joined sequentially by an recurrent encoder (e.g. RNN, LSTM, attention, etc.), and the resulted hidden state of the bottleneck layer was extracted as the learned features. It was then added Gaussian noise and copied to the input of a symmetric recurrent decoder, unrolled to the former dimensions, and then decompressed to the input dimensions via a stack of static layers symmetric to the static encoder. Under the semi-supervised setting, the learned features also took a parallel path of an NN-based regression model to predict the mood, health, and stress scores of the next day. Each component will be detailed in the following paragraphs.

**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 adopted two types of hierarchical local connections as follows, namely the Locally Connected MLP and 1D CNN.

1) **Locally Connected MLP** is described by Equation (3.1),

\[
a^l_j = \sigma \left( \sum_{i \in F^l_t} w^l_{ij} a^{l-1}_i + b^l_j \right)
\]

where in layer \( l \), \( a^l_j \) denotes the \( j \)-th node; \( F^l_t \) is the index set of nodes belonging to time frame \( t \); \( w_* \) and \( b_* \) are corresponding trainable parameters (weight and bias). The Rectifier Linear Unit (ReLU) is used as the activation function \( \sigma(\cdot) \).
With the high-resolution sensor data as the input, the static MLP encoder starts with a locally connected linear 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 $1440d_1 \rightarrow [L_{l2}] \rightarrow 8d_2 \rightarrow [L_{l3}] \rightarrow 8d_3$ is completed in multiple layers with ReLU activation stacked sequentially. The final dimensionality of the learned features is $8d_3$.

2) 1D CNN definition is given by Equation 3.2,

$$a_j^l = P\left(\sigma\left(\sum_{i=1}^{N} k_{ij}^l \ast a_{i}^{l-1} + b_j^l\right)\right)$$

(3.2)

where $k$ is the convolution kernel and $b$ is the corresponding bias; $j$ is the channel number out of $N$ total channels in the $(l-1)$-th layer; $P(\cdot)$ is the pooling layer that reduces the length of hidden nodes; $\ast$ is the convolution operator.

Different from locally connected MLP, the 1D CNN is by nature a local operation, because the aggregation of 1D CNN is a shift-and-sum operation with a trainable kernel sliding through the entire sequence. Also, stacking multiple convolution-activation-pooling blocks is by nature hierarchical. Thus, it saves us the need of explicitly defining ‘timeframes’. Moreover, CNN is translation invariant, meaning that it can capture featured local structures of the data regardless of their position in the data [92]. Combined with attention mechanism, 1D CNN has desirable properties to
identify important input/feature segments, thus helping deepening our understanding of the model’s behaviors.

**Static Autoencoder**

Autoencoders are essentially unsupervised neural networks, originally developed to learn efficient data representation, typically for anomaly detection [93], smart imputation [94] and nonlinear dimensionality reduction [95]. 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 MLP 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 weights can act similarly as regularization. Equation 3.3 gives the form of a single-layer tied
weight autoencoder,

\[ f_\theta(x) = \sigma_2(b_2 + W_{dec} \sigma_1(b_1 + W_{enc}x)) \],

where \( W_{dec} = W_{enc}^T \). \hspace{1cm} (3.3)

In Equation 3.3, \( b_1 \) and \( b_2 \) respectively denote the bias terms of encoder and decoder; \( \sigma_1 \) and \( \sigma_2 \) 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 \( \sigma_1 \) and \( \sigma_2 \), it would be equivalent to the 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,

\[ \hat{\theta} = \arg \min_\theta \frac{1}{2m} \left[ \sum_{i=1}^{m} (f_\theta(x^{(i)}) - x^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_{ij}^2 \right]. \hspace{1cm} (3.4) \]

where \( \theta \) denotes the to-be-optimized autoencoder parameters, and \( f_\theta(\cdot) \) is the encoding-decoding transformation; input raw data \( x \) is of length \( m \); \( \lambda \) 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 [96, 97, 98]. 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 our representation learning framework, fed on
intermediate static features at each time step. Its form can be simplified as follows,

Encoder: \( h_t = f_\phi(Uh_{t-1} + Wf_{\text{static}}(x_t)) \)

Decoder: \( h_t = f_\psi(Uh_{t-1}) \)

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. \( f_\phi, f_\psi, \) and \( f_{\text{static}} \) abstract other internal transforms (gates mechanism, activation functions, stacked layers, etc.) in LSTM encoder, LSTM decoder and static encoder.

### 3.2.3 Modality Fusion

In multi-modal machine learning problems, when and how to merge modalities remain an application-dependent choice. In this section, we compared two modality fusion schemes (Figure 3.3). We empirically chose 48 as the final feature dimensionality for both schemes. Features were learned via unsupervised representation learning frameworks.

The late modality fusion scheme, or late fusion, is illustrated by Figure 3.3a. SC, ST, and AC channels are propagated as completely separate data streams in the autoencoder. The computational paths are equivalent to training three independent representation models, one for each channel. Consequently, SC, ST, and AC unimodal features are distinguishable in the late fusion features.

The early modality fusion scheme, or early fusion, is illustrated by Figure 3.3b. SC, ST, and AC data are concatenated before being sent to the autoencoder. Each locally connected unit transforms data from all three channels through hidden layers into fused features. As a result, the early features are a mixture of multi-modal characteristics.
3.2.4 Semi-supervised Learning

In this thesis, the semi-supervised learning refers to the learning process where the total loss is the summation of a supervised portion and an unsupervised portion, as described by Figure 3.2. Both unsupervised and supervised losses are MSE loss, because both the sensor data and the wellbeing scores have been normalized to continuous values ranged 0–1 beforehand. The mathematical form is given by Equation 3.6,

\[
L_{\text{total}}(X, y|\theta) = \lambda L_{\text{recon}}(X|\theta_u) + (1 - \lambda)L_{\text{infer}}(X, y|\theta_s)
\]  

(3.6)

where \(L_{\text{recon}}\) represents the unsupervised reconstruction loss, and \(L_{\text{infer}}\) represents the supervised inference loss; \(\lambda\) is the balance factor which needs to be configured. During the training iterations, the system parameters \(\theta = \{\theta_u, \theta_s\}\) were optimized by backpropogating the \(L_{\text{total}}\) via stochastic gradient descent on mini-batched inputs of
the raw multi-modal sensor data.

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 variation of RNN that is well-suited for processing long time series with arbitrary lags 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 [20, 99, 98].

In an LSTM unit, cell remembers values over arbitrary time intervals, whereas input gate, output gate, and forget gate regulate the flow of information into and out of the cell [100]. As discussed by multiple studies [101, 102], stacked LSTM can learn at varied timescale across levels and thus incorporate richer temporal information than single-layer LSTM. 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. The dropout rate was set at 0.3 in all LSTM layers and 0.5 in the dense layers. Sequence-wise batch normalization was also applied, as suggested for a speech recognition problem where Amodei et al. [103] 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 3.7

\[
h_t^l = \sigma \left( \eta \frac{W_t^l h_t^{l-1} - E[W_t^l h_t^{l-1}]}{\sqrt{\text{Var}[W_t^l h_t^{l-1}]} + \epsilon} + \beta + U_t^l h_{t-1}^{l-1} \right)
\]

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

In terms of the objective function, since this is still a regression problem for continuous wellbeing scores, we are going to stick with the MSE loss as the wellbeing prediction loss \( L_{WP} \) whose form is given by the following equation,

\[
L_{WP}(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N} (y_i - \hat{y}_i)^2
\]

(3.8)

where \( y_i \) and \( \hat{y}_i \) are respectively the ground truth score and the model predicted score; \( N \) is the number of observations in a batch.

3.3.2 Personalized Model – Multi-task Linear Regularized Models

Multi-tasking learning (MTL) algorithm can optimize multiple different yet related tasks together \[104\] 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 (i.e. no task-specific information) and single-task learning (i.e. no shared information) \[105\].

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 study, we compared between MTL predictions with individual participants as tasks and with clusters of participants as tasks.

Alternatively, the individual-as-task MTL can be viewed as a special case of the cluster-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 [106] for clustering datasets with mixed numeric and categorical values. The number of clusters was determined at the highest mean Silhouette score of all samples [107]. 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 adopted the $\ell_{2,1}$ regularized MTL linear regression model whose objective function is as follows,

$$
\hat{\theta} = \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} ||\theta_T X_i - Y_i||_2^2 + \lambda_{2,1}||\theta||_{2,1} + \lambda_2||\theta||_2^2
$$

(3.9)

where $X_i$ represents the input matrix of the $i$-th task, and $Y_i$ is the label of samples belonging to that task; $\theta$ is the weight matrix. $\lambda_2$ controls the $\ell_2$-norm penalty; $\lambda_{2,1}$ controls the $\ell_{2,1}$-norm penalty.

### 3.3.3 Evaluation Metrics

**Mean Squared Errors (MSE)**

We evaluate the quality of the learned features by the MSE of the reconstructed and original input data. It measures the unsupervised learning performance of how close the original input can be recovered from the learned features. Lower values of MSE indicates better feature learning performance. The formula is given by Equation 3.10.

$$
L_{\text{MSE}}(X, \hat{X}) = \frac{1}{M} \sum_{i=0}^{M} (X_i - \hat{X}_i)^2
$$

(3.10)

where $X$ and $\hat{X}_i$ are respectively the original input data and the reconstructed output data; $M$ is the number of all observations to be evaluated by this metric.
Mean Absolute Errors (MAE)

One of the evaluation metrics of the learned features’ prediction efficacy is by the MAE of the predicted and true mood, health, and stress scores. It measures the supervised learning performance of how precise the wellbeing scores can be fit from the learned features. Lower values of MAE indicates better wellbeing prediction performance. The formula is given by Equation 3.11.

$$L_{MAE}(y, \hat{y}) = \frac{1}{M} \sum_{i=0}^{M} |y_i - \hat{y}_i|$$

(3.11)

Transition Accuracy

In addition to the absolute precision our prediction models can achieve for each individual day, we care about whether the models can predict the change in score instead of the raw score? We believe that a user also wants to know this information – is tomorrow going to be better or worse? Therefore, transition accuracy (TA) is introduced as given by the following Equations 3.12–3.15, which were adapted from [108].

$$\Delta y_t = y_{t+1} - y_t$$

(3.12)

$$\Delta \hat{y}_t = \hat{y}_{t+1} - y_t$$

(3.13)

$$Error_t = \begin{cases} 
0 & \text{if } sign(\Delta y_t) \cdot \Delta \hat{y}_t > 0 \\
1 & \text{otherwise}
\end{cases}$$

(3.14)

$$TA = \frac{1}{T} \sum_{t=1}^{T} (1 - Error_t)$$

(3.15)

where $y_t$ denotes the self-reported wellbeing score(s) at day $t$, with $\hat{y}_t$ being the predicted score(s) of that day. $\Delta y_t$ is the true transition of a user’s self-reported wellbeing from day $t$ to day $(t+1)$; $\Delta \hat{y}_t$ is the predicted transition of a user’s wellbeing from day $t$‘s self-report to day $(t + 1)$’s prediction. True transition directions of two
consecutive days were filtered at a threshold $|dy|$; that is to say, any $\Delta y_t$ instances such that $|\Delta y_t| < |dy|$ were zeroed out. $T$ is the number of remaining true transitions (positive or negative; zero transitions were discarded).

### 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. [109] 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 3.4 [110]. The
softmax-normalized attention weights quantify the importance scores of a sequence.

### 3.4.2 Correlation Analysis

Correlation analysis is used to evaluate the strength of a relationship between two variables under certain assumptions \[111\]. 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 features that are important to personalization performance, we look for features that 1) contribute the most in a task or 2) result 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 cluster 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. For each feature, if 1) its clustered weight coefficients have statistically different distributions, and 2) opposite signs are observed in clusters’ means, it would be considered as critical to personalization, or a critical feature.
Chapter 4

Experiments

Our general goal is to use deep learning methods to automatically learn and extract efficient and interpretable features from high-resolution wearable sensor data. 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. To achieve that, a series of experiments were designed and completed including

1. Unsupervised representation learning of wellbeing features (Section 4.1),
2. Semi-supervised representation learning of wellbeing features (Section 4.4),
3. Semi-supervised learning of visual attention on input data (Section 4.4.1),
4. Generalized and personalized wellbeing prediction (Section 4.2),
5. Interpretation of the auto features and wellbeing predictions (Section 4.5).

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 1.5. The linear MTL prediction models were adapted from the MALSAR toolbox which was proposed by Zhou et al. [112] and available via http://jiayuzhou.github.io/MALSAR/
4.1 Unsupervised Representation Learning

In the first part of experiments, which built up the main body of this thesis, we focused on the unsupervised representation learning. In the following context, we would be referring to the unsupervised framework if the supervision strategy is not explicitly named. We are particularly interested in the unsupervised representation learning mode because it does not require labeled data. In recent health-related studies, raw data are becoming increasingly accessible thanks to automatic ubiquitous devices, but high-quality labels are still costly and rare where human labor is to be involved. Thus, we started with the unsupervised representation learning, with a goal to verify that we can achieve good representation learning performance and wellbeing prediction performance without ground truth. If found true, this will greatly loosen the heavy burden of label acquisition and benefit researchers and groups with similar interest. Technically, the unsupervised architecture of representation model is simpler than the semi-supervised one, encompassing less sources of uncertainty yet greater controllability when studying configurations of the model.

In terms of the hyperparameters, the representation learning model was optimized via the Adam algorithm, with the learning rate being 0.003 and trained for 100 epochs. In the Adam optimizer, $\beta_1$ and $\beta_2$ were set to 0.9 and 0.999 respectively, and the weight decay was $10^{-6}$. Our prior work [113] 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.

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.

4.1.1 Input resolution

Although high-resolution data were available, storing, processing, or transmitting these raw data may generate extra cost. Thus, we attempted to reduce the sampling rate of sensors and look into the representation learning performance of lower resolutions. More specifically, we compared the reconstruction loss (MSE) using 8, 4, 2, and 1 Hz raw sensor data input to the representation model. The lower-resolution data were synthesized by downsampling 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

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 [114, 115, 116] 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 this study, we are also interested in the minimum acceptable size of training set for time/energy saving purpose. Therefore, we probed into the impact of training data amount on the reconstruction loss for the unsupervised representation learning model by constructing different sized subsets for training. All samples were selected at random. The baseline amount was set to be 60% of the entire dataset, for approximately 3900 days in total, or 18 days per participant on average. 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 3.2).

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 representation model under the user-dependent assumption. The trained general representation model 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 representation 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 with Personalization Policies

Two schemes were designed to split 60% train, 20% validation, and 20% test data for user-dependent and user-independent personalization policies. 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

Wellbeing prediction can take the form of a generalized LSTM model (Section 3.3.1) or a personalized MTL model (Section 3.3.2). We had different models for generalized and personalized prediction, because those two models were the best performing models corresponding to the two personalization assumptions. In other words, in the generalized scenario, the LSTM prediction model yielded better prediction performance, whereas in the personalized scenario, the linear MTL model generated the most precise future wellbeing scores. Our experiments were designed for both forms with a highlight on the personalized models because it can result in higher prediction performance while being able to adapt to new users via a relaxation on the
personalization policy. 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 $\lambda_2$ 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. The crafted features were in 34 unique feature definitions; they were computed from four non-overlapping time frames, namely 0-3H, 3-10H, 10-17H, and 17H+, resulting in 136 crafted features daily. Pearson correlation was then computed between the learned features and the crafted features. By ranking the statistically significant correlations, we could depict some intuition of the learned features and prediction behaviors.

### 4.2.2 Number of Previous Days

Previous studies [8, 20] 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 Policies

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.

We committed an exhaustive search for the number of groups from 2 to 147. Silhouette score constantly increased as we had more groups, reaching the highest value of 0.60 at \( k = 147 \), equal to the number of unique participants in the training set. In the case of predicting an unseen user’s wellbeing, we firstly applied the same clustering algorithm to his or her gender and personality, locating him or her at one group with the closest distance. Then, we adopted the trained weights of this group for the target user’s sensor features and made predictions accordingly.

4.3 Modality Fusion

In this last section, we compared different modality fusion schemes with regard to the personalized wellbeing prediction (detailed in Section 4.1.4). We compared among all combinations of three modalities (SC, ST, AC, SC+ST, SC+AC, ST+AC, SC+ST+AC). Experimental settings (e.g. number of layers, hidden node size, input data resolution and amount, etc.) were strictly controlled in order to produce con-
vincing comparison result. Only the learning rate was tuned to secure appropriate gradient descent during training with different input data and combinations. Following the representation model, personalized wellbeing prediction performance was compared between modality fusion schemes, as described in Section 4.2.3. Additionally, to understand the sources of difference between prediction behaviors based on the late features and the early features, we exploited the critical feature analysis as Section 4.5.2 described. With that said, if not clearly stated, the default configuration is trimodal late fusion throughout this thesis.

4.4 Semi-Supervised Representation Learning

Unsupervised representation learning is particularly favorable for small or unlabeled datasets. In this thesis, a well-labeled dataset was used, so we were also interested in whether ground truth labels might help with the representation learning process as well. Therefore, extra experiments were designed to enable supervision within our representation framework, and its results were to be compared with the unsupervised settings in terms of both generalized and personalized wellbeing prediction, feature correlation analysis, and the MTL model weight analysis.

With similar experiment configurations as listed in Section 4.1, we took a step further and integrated the inference module to the unsupervised representation model to construct a semi-supervised representation learning model as depicted by Figure 3.2. As a pilot study, in order to focus on the comparison between supervision strategies, we controlled the other configurations of our framework. The base representation model was a locally connected LSTM denoising autoencoder that took 8 Hz multi-model data with early fusion. It was trained also with self-reported mood, health, and stress scores in a non-personalized manner; the scores were multiplied
by a normalizing factor of 0.01. Although the semi-supervised representation model alone was a non-personalized model (i.e. trained without considering the identity of data providers), we still needed to carefully choose a data split policy as discussed in Section 4.1.5 in order to be consistent with the following prediction model’s personalization strategy. To be specific, we focused on the user-dependent assumption and applied individual-as-task personalization policy for the personalized wellbeing prediction model. Thus, the data preparation for semi-supervised representation learning followed the user-dependent split scheme. The data split approach was kept the same for the generalized wellbeing prediction for consistency, though technically a generalized prediction model does not care about whether all users have been seen during training. We will show the training curves of the reconstruction loss and the inference loss separately. With the semi-supervised auto features, we revisited Section 4.2 to verify its wellbeing prediction performance.

4.4.1 Convolutional Visual Attention

In this section, we modified the architecture of both static and temporal layers of the semi-supervised representation learning framework, in order to learn high-resolution visual attention to gain insights of the relation between the sensor data and wellbeing status. The convolutional visual attention was learned via a semi-supervised temporal representation model where the static encoder was three stacked blocks of 1D CNN, nonlinear activation, index-preserving max pooling, and batch normalization layers; the temporal encoder and decoder were multihead self-attention layers; the static decoder was symmetric to the static encoder with max unpooling and tied-weight transposed 1D CNN layers. The input data was 8 Hz raw SC, ST, and AC sensor data, and for the ground truth labels we used mood, health, and stress scores. The learning
rate was empirically set at a relatively small value of $5 \times 10^{-5}$ to produce the best observable gradient descent. In addition to the learning curves of the convolutional visual attention, we will also unveil the upsampling process from 100 base attention weights to 86,280 high-resolution attention map.

Since SC reflects only sympathetic activities of the autonomic nervous system, SC has been recognized as a classical biomarker of cognitive stress in psycho-physiological studies [81, 82]. Consequently, we are especially interested in investigating the relationship between stress and SC data. Thus, after obtaining the high-resolution attention map, we applied the prominence peak detection algorithm to the SC attention map to look for peak attention moments within the SC data. The input segments corresponding to the high-attention moments are recognized as important physiological markers of the cognitive stress. Due to the uniqueness of this task and the nature of deep interpretation, there is hardly a universal or quantitative measure of the goodness of visual interpretations. That said, we provide real data examples randomly drawn our test set.

4.5 Multi-Tiered Interpretation

4.5.1 Learned-Crafted Feature Correlation Analysis

To interpret the learned features, we employed 34 crafted features of the raw sensor data as mentioned in Section 4.2.1. To explain what exactly did the representation model learned (i.e. what the auto features meant), we investigated the significant Pearson correlations between the learned features and the crafted features. The higher a correlation is, the more similar a pair of learned and crafted features are to be identified. It could also be interpreted as that the learned feature contained some
information about the corresponding crafted feature, thus probing into the ‘meaning’ of certain learned features.

4.5.2 Critical Feature Analysis by MTL Weights Clustering

Additionally, 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 Section 4.2.3. For each feature, if its weight coefficients had statistically different distributions across different clusters, and especially if opposite signs were observed in cluster means, it would be considered as critical to personalization, or a critical feature.

4.6 Statistical Test

Analysis of variance (ANOVA) test was used in testing differences among more than two groups of data. Following ANOVA, Tukey HSD test 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, semi-supervised auto features, etc.).
Chapter 5

Results

5.1 Unsupervised Framework Based Results

5.1.1 High vs Low Input Resolution for Learning Auto-features

Figure 5.1 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.

5.1.2 Amount of Training Data for Learning Auto-features

Figure 5.2 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.
Figure 5.1: Reconstruction loss on validation set with different input data resolution for each channel.

Therefore, we can safely set the focus on the comparison of the training set size itself.

5.1.3 Temporal vs Static Autoencoder Architecture

Figure 5.3 compares the curves of reconstruction loss on the user-dependent validation set with different autoencoder architectures as described in Section 4.1.3. It emphasizes the advantage of the temporal component – the LSTM encoder-decoder wrapper – in the representation model. Although both models started off similarly, without the temporal component, the static autoencoder could not reach a local optimum as far as the temporal autoencoder. Here we again combined channel-wise losses because all channels shared a consistent tendency.

5.1.4 Personalizing the User-Dependent Representation Model

Figure 5.4 shows the histogram of changes in participant-specific MSE by fine-tuning the general representation model to each individual. Given a participant, a nega-
ative 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% of the participants. In terms of the failure cases, possible reasons may include overfitting, simpler underlying structure of data or the limited amount of finetuning samples.

Figure 5.2: Reconstruction loss vs training-set size.

Figure 5.3: Reconstruction loss vs autoencoder architecture.

Figure 5.4: 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.
5.1.5 Auto-learned vs Hand-crafted Features: Prediction Performance Comparison

To demonstrate that our temporal representation model can learn efficient features to predict multiple wellbeing labels, in Table 5.1 we compared its performance with benchmark results derived from prior works on the same topic [113, 20]. 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 5.1, 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 temporal representation model, 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.

5.1.6 Generalized Multi-day LSTM Wellbeing Models

Figure 5.5 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 predic-
Table 5.1: Wellbeing prediction mean absolute errors (MAE) using individual-as-task MTL on different feature sets (Mean±S.D.).

<table>
<thead>
<tr>
<th>Label</th>
<th>Crafted Features</th>
<th>Auto Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Static</td>
<td>Temporal</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mood</td>
<td>16.4±0.3</td>
<td>14.6±0.3</td>
</tr>
<tr>
<td>Health</td>
<td>15.8±0.3</td>
<td>14.3±0.5</td>
</tr>
<tr>
<td>Stress</td>
<td>16.7±0.3</td>
<td>15.7±0.3</td>
</tr>
</tbody>
</table>

* 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).

Nevertheless, we found that temporal features showed a significant advantage over static features only on mood and stress using 7-day data.

Figure 5.6 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.

5.1.7 Clustering Based on Profile

We committed an exhaustive search for the number of groups from 2 to 147. Silhouette score constantly increased as we had more groups, reaching the highest value
Figure 5.5: 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.

Figure 5.6: Attention weights distribution on time steps in LSTM prediction model using 7-day temporal features.

of 0.60 at $k = 147$, equal to the number of unique participants in the training set. Figure 5.7 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.

5.1.8 User Dependent vs Independent Policies

Figure 5.12 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.

In the user-dependent case (Figure 5.12b), 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 $14.1 \pm 0.2$, $12.4 \pm 0.3$, and $15.0 \pm 0.3$.

In the user-independent case (Figure 5.8b), 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 5.1.7. The lowest achieved prediction MAE values were $14.5 \pm 0.4$ for mood, $14.4 \pm 0.4$ for health, and $15.7 \pm 0.5$ for stress.

5.1.9 Interpretation of Features and Predictions

Correlation Analysis

We computed the correlation matrix between the corresponding modalities of auto-learned features and hand-crafted features. The significance-filtered ($p < 2.3 \times 10^{-5}$, the adjusted p-values [120]) correlation heatmap for daily features is shown in Figure 5.9. 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).

In Table 5.2 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 fea-
Figure 5.9: 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 $\pm 0.05$ ($p < 2.3 \times 10^{-5}$). The auto SC, ST and AC features were independently learned from the corresponding modalities and were not personalized.

Table 5.2: Most interpreted auto features for each sensor modality, as well as the top-5 highest correlated crafted features with associated correlation coefficients ($p < 2.3 \times 10^{-5}$).

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

According to Table 5.2, 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 5.10, 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).

![Figure 5.10: 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.](image)

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).

5.2 Late vs Early Modality Fusion

5.2.1 Representation Learning

When to fuse different modalities has an impact on the performance of representation learning. We compared the reconstruction MSE loss via late and early fusion of modalities during the representation learning process. According to our observation from Figure 5.11, early modality fusion could provide a slightly steeper and deeper learning curve than late fusion. Despite a similar trend, the early fusion always produced lower reconstruction loss than the late fusion. The early fusion loss was eventually reduced from 0.40 (S.D. = 4.4×10⁻³) to 0.034 (2.1×10⁻³), while the late fusion loss concluded at 0.059 (6.3×10⁻³).

It should be noted that the late fusion curve by Section 3.2.3 was derived using a unified training scheme for all modalities, but late fusion can be potentially more flexible. With late fusion, it would be much easier and effortless to use different train-
5.2.2 Wellbeing Prediction

Moving forward to predicting wellbeing scores, a coincident pattern was observed – for all labels, combinations and both personalization settings, the early fusion features always resulted in statistically equivalent or significantly lower errors compared to corresponding late fusion features, as Figure 5.12 shows. Using trimodal data, when switching from late to early fusion features, the averages of prediction MAE on unseen participants dropped by 4.8%, 6.1%, and 2.4%, landing at 15.8 (S.D. = 0.4), 15.4 (0.3), and 16.5 (0.2), respectively for predicting mood, health, and stress (p<0.05). Overall, the early fusion of trimodal data resulted in the extraction of more robust and predictive deep sensor features.

Figure 5.11: Comparison of the unsupervised representation learning loss on the timing of fusing three modalities (skin conductance, skin temperature, and acceleration). The losses were computed on the validation set during a iterative mini-batch training session.
Figure 5.12: Performance of wellbeing prediction using user-dependent and user-independent personalizing MTL on different fusion schemes and modality combinations. For the fused modalities, the inner bars denote early fusion; the outer bars denote late fusion.

5.2.3 Critical Feature Analysis

We conducted the feature and prediction analysis to our best performing trimodal health model (user-dependent), as described in Section 4.5.2. Taking the health model for instance, we found that two clusters resulted in the highest Silhouette scores, namely 0.59 for the late fusion weight coefficients and 0.58 for the early. Four early fusion features were found critical to personalization, whereas only one late fusion feature was critical. This observation indicated that early features might contain more crucial and personalized information than late features.
5.3 Semi-Supervised Framework Based Results

From our experience working with the unsupervised representation framework in Section 5.1 as well as on similar topics (e.g. unsupervised autoencoder based representation learning for sleep detection), we found that the dimensionality of the bottleneck layer (i.e. the number of auto features) did not much influence the reconstruction loss, nor was it a booster of the prediction performance. Previously, we tested 2, 4, 6, and 12 as the feature dimensions of each timestep and came to this observation. The total feature dimensionality should be a multiplication of per-timestep dimensionality by eight then by three, given that we defined the eight 3-hour timeframes and took three channel multi-modal data as the input as Figure 3.2 illustrated.

However, this did not hold true for the semi-supervised models. The introduction of the MLP inference module made the model, especially the reconstruction module, difficult to converge. Challenges came from the mismatch of the learning capacity of the reconstruction and the inference modules, since the reconstruction module encompassed much more parameters than the inference module. It took us great efforts while trying out all possible tweaks including a different balance factor $\lambda$ in the semi-supervised loss, applying different learning rates or even different optimizers to the reconstruction and inference modules, etc. Unfortunately, we could only solve the problem with the timestep-wise feature size greater or equal to 4 (i.e. total feature dimensions = 96), which was twice as much as the unsupervised case.

As a result, we have to report our results based on 96 semi-supervised auto features. In order for fair comparison, we reran a subset of the unsupervised experiments to obtain 96-dim unsupervised features as the benchmark results for this section.
5.3.1 Representation Learning

The semi-supervised representation framework consists of an unsupervised locally connected recurrent autoencoder as the reconstruction module and a supervised MLP as the inference module. Respectively, the reconstruction module should learn some compressed general information mostly useful for recovering the sensor data itself from the low-dimension feature space to the original data space; the inference module should learn information predictive of the future wellbeing status. This framework was trained in a non-personalized manner, i.e. it resulted in universal parameters for all participants without discriminating between identities.

Figure 5.13 shows the total loss as the weighted sum of reconstruction loss and inference loss. All three types of losses were generally reduced on the validation set during the training process, indicating that both the reconstruction and inference modules were able to learn useful information from the raw multimodal sensor data.

Figure 5.13 : Semi-supervised learning curves of weighted inference loss, weighted reconstruction loss, and total loss on the validation set. Different y-axes were used to plot reconstruction loss (right vertical axis) and inference loss (left vertical axis) in order for better illustration. The total loss is equal to the sum of the weighted reconstruction and inference losses ($\lambda = 0.5$).
Given that an inference module has been inserted to form the semi-supervised framework, it also introduced an external source of loss and a new gamer in the optimization process. Thus, we believe that directly comparing between the losses of the semi-supervised and unsupervised representation models would not be convincing enough to conclude their pros and cons. Therefore, in the following part, namely Section 5.3.2, we compare the two supervision schemes on a higher level – we look at the wellbeing prediction performance produced by several settings and models resulted from the two supervision schemes.

### 5.3.2 Wellbeing Prediction

The wellbeing prediction models are essentially no different from the ones that used unsupervised features in Section 5.1.5 and Section 5.1.6. Here, we compared the semi-supervised vs unsupervised features based prediction with three prediction models, 1) generalized one-size-fits-all LSTM on 7-day features; 2) user-dependent individual-as-task MTL on 1-day features; and 3) user-dependent individual-as-task MTL on 7-day features.

#### Semi-Supervised vs Unsupervised Comparison

In Table 5.3 we compiled all prediction performance of these models. From these results, we can easily observe that with the generalized 7-day LSTM prediction model, the semi-supervised features were more powerful than the unsupervised features in prediction future health and especially stress; for the other label, however, learning features from raw sensor data with ground truth scores did not seem to boost its predictive efficacy of the participants’ mood. With regard to the personalized MTL models, the advantage of supervision appeared even dimmer in the 1-day MTL predic-
tion model, and it completely vanished as we incorporated more history information (7 days’ features) to the MTL prediction model. Overall, we did not observe a revolutionary improvement by incorporating ground truth information in the representation learning stage.

Table 5.3: Wellbeing prediction performance MAE (S.D.) by the automatically learned features via semi-supervised vs unsupervised representation learning. An asterisk indicates a statistically significant advantage in performance (p<0.05).

<table>
<thead>
<tr>
<th>Prediction Model</th>
<th>Representation Model</th>
<th>Semi-Supervised Auto Features</th>
<th>Unsupervised Auto Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Generalized</td>
<td>Mood</td>
<td>18.5 (0.4)</td>
</tr>
<tr>
<td></td>
<td>LSTM (7 Days)</td>
<td>Health</td>
<td>19.1 (0.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stress</td>
<td>19.0 (0.4)*</td>
</tr>
<tr>
<td>Personalized</td>
<td>1 Day</td>
<td>Mood</td>
<td>14.1 (0.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Health</td>
<td>12.4 (0.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stress</td>
<td>15.0 (0.3)</td>
</tr>
<tr>
<td></td>
<td>MTL</td>
<td>Mood</td>
<td>14.1 (0.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Health</td>
<td>12.4 (0.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stress</td>
<td>15.0 (0.3)</td>
</tr>
</tbody>
</table>

Personalized Prediction Performance vs Self-Report Variability

From our experiment, we noticed that the prediction model’s performance varied from user to user. For some users, the model could make very precise prediction, whereas on some other users, the model could not always make accurate prediction of his or her wellbeing scores. In order to provide an intuitive view of the source of this inter-user performance gap, in Figure 5.14 we plotted each participant’s health prediction
Figure 5.14: Personalized health prediction performance could be strongly impacted by the variability in ground-truth health scores. Higher prediction performance was achieved with more stable participants.

MAE against his or her self-reported scores S.D. We found a good fit between the two statistics, and thus we could confirm that the instability of ground truth was a barrier to accurately predicting future wellbeing.

**Transition Accuracy of the Predicted Scores**

Table 5.4 shows the transition accuracy evaluations of health prediction performance of the semi-supervised features. The baseline accuracy values were obtained by taking the majority guesses within each individual’s transition ground truth set. In most cases, our health transition predictions were significantly better than the baseline (the same applied to mood and stress as well). Especially, for very large swings $|dy| = 40$, ...
the personalized health model could correctly forecast every single transition with 100% accuracy.

Table 5.4: Accuracy of tomorrow’s health forecast for a change of self-reported sick-healthy scores that is at least \(|dy|\) different tomorrow compared to today, based on generalized and personalized prediction models with 7-day features. Asterisks denote how significantly that our predicted transition accuracy outperformed the corresponding baseline result by McNemar Test (* \(p < 0.05\); ** \(p < 0.01\); *** \(p < 0.005\); **** \(p < 0.001\)).

| \(|dy|\) | Baseline | Generalized LSTM | Personalized MTL |
|--------|----------|------------------|------------------|
| 1      | 52.2% (71/136) | 66.2% (90/136) ** | 69.1% (94/136) ** |
| 5      | 54.9% (56/102) | 68.6% (70/102) *  | 76.5% (78/102) *** |
| 10     | 55.4% (41/74)  | 74.3% (55/74) **  | 83.8% (62/74) **** |
| 20     | 50.0% (20/40)  | 75.0% (30/40)     | 90.0% (36/40) *** |
| 40     | 59.1% (13/22)  | 90.9% (20/22) *   | 100.0% (22/22) *** |

5.3.3 Statistical Interpretation

In this section, we repeat the statistical interpretation for the semi-supervised features. Again, the principles remain exactly the same as that of the unsupervised features in Section 5.1.9. We compare the statistical patterns between the semi-supervised and unsupervised features’ results, in search for how differently the representation model might have incorporated certain information and how that different information might influence the predictive efficacy of the learned features.
Feature Correlation Analysis

We provide the crafted vs auto features correlation map in Figure which was organized following the same principle of Figure 5.9. That said, it should be noted that here we plot the correlation map between 96 auto features and 136 crafted features. Now that the number of auto features doubled, we should lower the significance threshold accordingly to $1.7 \times 10^{-5}$.

![Correlation Matrix](image)

**Figure 5.15**: 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 $\pm 0.05$ ($p < 1.7 \times 10^{-5}$). The auto SC, ST and AC features were independently learned from the corresponding modalities and were not personalized.
More specifically, Figure 5.15a displays the correlation map that involves 96 unsupervised auto features (32 auto feature per modality), and Figure 5.15b displays that of the 96 semi-supervised auto features (32 auto feature per modality). In Figure 5.15a, the ranks of significant correlation matrices were 31, 20, and 7, respectively for crafted features’ correlations with SC, ST and AC features. In Figure 5.15b, the ranks were 23, 24, and 11. Generally speaking, a higher rank indicates more significantly correlated items. Also from visual comparison between these two groups of matrices, it can be observed that the unsupervised features were better represented by the SC crafted features (e.g. 17H+: Mean Unnorm, 17H+: Median Norm, 17H+: Median Unnorm, etc.), whereas the semi-supervised auto features were better described by the ST (e.g. 17H+: Median Raw, 17H+: Median Stillness, 10H-17H: Minimum Raw, etc.) and AC (e.g. 3H-10H: Stillness Percent, 3H-10H: Step Count, 17H+: Stillness Percent, etc.) crafted features. Even though, readers should be advised that the idea of going deep to learn features from raw data was motivated by the hope that a data-driven representation learning model may be able to learn beyond human expert knowledge. Therefore, as a measure based on a finite set of expert-defined physiological and behavioral features, this correlation analysis merely aimed to provide an entry point of possibly interpreting the deeply learned features, yet it does not necessarily indicate the prediction efficacy or quality of any feature set. Thus, this result does not contradict with the wellbeing prediction comparison from Section 5.3.2.

Individual-based Weights Analysis for MTL Wellbeing Prediction

Just as what we did in Section 5.1.9, here we cluster the individual task weights of the health prediction MTL model using the semi-supervised late fusion features. Four clusters gave the highest Silhouette score of 0.53. The four clusters each contained
172, 26, 22, and 19 individual tasks. Among the 96 semi-supervised auto features (32 features per modality), we identified 7 critical SC features, 10 critical ST features, and 11 critical AC features. The top correlated crafted features included the following items (the highest correlation coefficient is also listed with each crafted feature),

- **SC:** 17H+: Percent Sensor Off (-0.07), 17H+: Median Norm (0.07), 17H+: SD Norm (0.07), etc.

- **ST:** 0H-3H: Median Raw (-0.31), 10H-17H: Median Stillness (-0.30), 3H-10H: Median Raw (-0.29), 0H-3H: Median Stillness (-0.28), 10H-17H: Median Raw (-0.28), etc.

- **AC:** 0H-3H: Stillness Percent (-0.13), 0H-3H: Step Count (0.08), 10H-17H: Mean Movement Step Time (0.06), etc.

Now we compute the benchmark result using unsupervised feature set, where two clusters gave the highest Silhouette score of 0.43. The two clusters each contained 182 and 56 individual tasks. Among the 96 unsupervised auto features (32 features per modality), we identified 7 critical SC features, one critical ST features, and zero critical AC features. The top correlated crafted features included the following items (the highest correlation coefficient is also listed with each crafted feature),

- **SC:** 17H+: Median Norm (0.25), 17H+: Mean Unnorm (0.24), 17H+: Median Unnorm (0.20), 17H+: Median Peaks 30min (0.18), etc.

- **ST:** 17H+: Median Stillness (-0.36), 17H+: Median Raw (-0.35), 10H-17H: Min Raw (0.21), 17H+: S.D. Raw (-0.16), and 10H-17H: S.D. Raw (-0.13).

- **AC:** None.
5.4 Semi-Supervised Convolutional Visual Attention Map

The semi-supervised convolutional visual attention (CVA) model consists of an unsupervised CNN-Attention autoencoder as the reconstruction module and a supervised MLP as the inference module. The outcome difference between a CVA model and a locally connected semi-supervised representation learning (SRL) model was the former had much higher-dimension latent space at the bottleneck layer than the latter. In this study, we specifically designed the CVA architecture so that its bottleneck layer had 100 “time-steps”, each containing 64 features, resulting in totally 6400 bottleneck features. Yet, in the SRL model, the number of time-steps was only 8, and the total number of features was 96. This difference was actually rooted in the purpose of these two models, where one should be reminded that the SRL model aimed to learn features which were suitable for predicting wellbeing scores (focus on prediction). Consequently, high feature dimensionality was not welcomed as it would cause a lot of problems including overfitting, curse of dimensionality, etc. On the contrary, the CVA model was aimed to associate the input data with a fine-grained attention map (focus on reconstruction), thus a reasonably adequate size of base attention weights was important to successfully construct a high-resolution CVA map. Not surprisingly, it came with a price that the wellbeing prediction performance was compromised.

Figure 5.16 shows the total loss as the weighted sum of reconstruction loss and inference loss in the semi-supervised CVA model. All three types of losses were generally reduced on the validation sets during the training process. This plot came from the SC channel, indicating that both the reconstruction and inference modules were able to learn useful information from the raw unimodal SC sensor data and the wellbeing scores.

Figure 5.17 shows an intermediate upsampling example outcome up to the first
Figure 5.16: Semi-supervised learning curves of the convolutional visual attention model based on the SC data and mood, health, and stress scores. The total loss is equal to the sum of weighted reconstruction and inference losses ($\lambda = 0.5$).

An unpooling layer following the bottleneck layer in the CVA model. The maximum upsampling rate shown on this plot was approximately x30. Compared to the base attention weights, both intermediate upsampling outcomes, with or without indices, exposed smoother attention distributions. Nonetheless, it is obvious that upsampling with the max-pooling indices resulted in more details in the high-resolution attention map than it without the pooling indices.

Figure 5.17: Segment of an upsampling example of attention weights by Gaussian RBF interpolation, with and without max pooling indices.

Figure 5.18 zoomed out and provided a view on the completed upsampling (rate
x862.8) of an entire day’s attention map, as well as a match between the base attention weights with the original input SC data.

Figure 5.18: Upsampling convolutional visual attention with max-unpooling. (a) displays the original input SC data in the top plot, and the bottom plot is the base attention weights; vertical dash lines separates the 100 base attention weights, and we call the position of each attention weight an attention moment. (b) shows the base attention weights as a line plot and a heatbar. (c) shows the completed upsampling of the plots of (b), resulted in 86,280 upsampled attention moments. The upsampling rate from (b) to (c) was x862.8.

With the upsampled attention map, we were then interested in identifying the moments when the CVA model paid high attention to the input data. To achieve that purpose, the prominence-based algorithm can detect good peaks and discard the noisy ones. Figure 5.19 shows the prominence (≈0.02) peak detection result of the
upsampled attention map from the Figure 5.18 instance. A total of 16 peaks were
detected and annotated on the plot.

Figure 5.19: An instance of the prominence peak detection results on an upsampled
attention map.

Figure 5.20 and Figure 5.21 show the original input data, base and upsampled
attention weights, and attention peaks for ST and AC data. They were all showing
the same instance as Figure 5.18 with subfigures organized using the same protocol.

Figure 5.20: Upsampling ST convolutional visual attention with max-unpooling.
Prominence peaks are annotated in (c).
Figure 5.21: Upsampling AC convolutional visual attention with max-unpooling. Prominence peaks are annotated in (c).

Identifying high-attention moments (HAM) enabled us to probe into the original sensing data. Take SC for example, we could extract data segments corresponding to the HAMs and look for patterns related with stress levels. Based on the participants’ self-report, we defined days with stress/calmness scores smaller than 33 as high-stress instances (binary label 1), and days with stress/calmness scores greater than 67 as low-stress instances (binary label 0). Accordingly, HAM data segments from a high-stress day would be labeled as high stress (binary label 1), and HAM data from a low-stress day would be labeled as low stress (binary label 0).

Figure 5.22 shows instances of HAM data segments randomly drawn from high-stress days and low-stress days. Although we managed to obtain attention moments with resolution as high as close to 1 Hz, here we plotted a 10-second neighborhood centered at the high-attention moments, so that the patterns of data at these moments can be more intuitively illustrated.
Figure 5.22: Ten-second SC data segments centered at high-attention moments.

(a) High-stress HAM data segments

(b) Low-stress HAM data segments
Chapter 6

Discussion

From our experiment results and analysis, we found that our proposed representation learning framework could efficiently learn temporal features from high-resolution sensor data, and those features were reliable estimators for forecasting one’s subjective wellbeing. In this section, we compare our results with prior work and outline the implications of our results for researchers in the 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 5.1.1 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 the temporal representation learning model 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. How-
ever, 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 5.3 that compared to static features, temporal features produced lower reconstruction loss during representation learning, but in Figure 5.5 we did not observe significant improvement using one-size-fits-all LSTM model to predict mood, health or stress. Similarly, when Jaques et al. [2] 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 [121], emotion recognition based on physiological signals [122], etc. Our proposed the temporal representation learning model 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 the temporal representation learning model can generalize with smaller partial training sets as well as the full set (Figure 5.2).

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 [46] 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 [46] 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. [123], improved their results of using several recurrent architectures on a text prediction dataset. Additionally, Laurent et al. [124] 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 distribu-
tion 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 When to Fuse Which Modalities

Revisiting Figure 5.12, when compared between different modality combinations, no significant difference was ever observed between bimodal and trimodal features, neither across three sets of unimodal features individually. Nonetheless, we found that unimodal features never outperformed multi-modal features. Furthermore, varied by fusion approach and target label, sometimes the prediction precision of multi-modal features significantly exceeded that of the unimodal features, and other times the gap was insignificant. For example, considering the user-dependent case, when predicting mood, multi-modal features significantly outperformed unimodal features, regardless of which fusion approach was applied. When predicting stress, neither late or early fusion showed a significant difference between multi-modal and unimodal features. As for the health prediction, however, the early fused multi-modal features significantly outperformed the unimodal features (p<0.05), while the late fused features did not. Considering the user-independent case, the patterns were interestingly consistent across all labels, where the early fused multi-modal features significantly outperformed the unimodal features (p<0.05), but the late fused multi-modal features did not.

We observed that the early and late fusion features always showed different predictive power of future mood and health. Based on our observation, we hypothesize
that early fusion outperformed late fusion in personalized tasks because it resulted in more sufficient capture of critical features. As another potential explanation, with early fusion, features can be smoother and more coherent than with the late fusion. According to [125] where the authors remarked a similar advantage of early over late modality fusion in image-based affect recognition, the reason could be that early fusion benefited from minor low-level architectural elements which were crucial to performance through deep propagation.

Furthermore, we anticipate that human health or other wellbeing is easier to infer with greater personalized concentration on physiological (ST and SC) information 24 hours prior. Our early fusion framework emphasized features about ST medians and lower-extremities, supporting current research (e.g. [126, 127]) that wrist ST can indicate thermal comfort which is closely related to perceived health. Additionally, low SC levels are associated with several health issues including chronic fatigue [128] and pain [129]. Prior work has also revealed strong links from social interaction to happiness [130], suggesting that exploiting phone and other ubiquitous data is very likely to boost our model’s performance.

6.3 The Role of Supervision in Representation Learning

The term *semi-supervised learning* is manifested in two seemingly opposite ways under different context. In studies aimed at supervised tasks (U→S), semi-supervised learning refers to augmenting a small set of labeled data with a large set of unlabeled data in order to solve a supervised learning approach. For example, Le et al. [131] investigated a linear autoencoder trained on joint reconstruction loss and inference loss, a similar design as our semi-supervised representation learning framework, where they found that the unsupervised auxiliary tasks could improve generalization perfor-
mance both theoretically and empirically. Recently, Latif et al. [132] tackled speech emotion recognition (SER) with multi-task semi-supervised learning framework to employ large available datasets with different labels, such as gender or speaker recognition, as auxiliary tasks. On the other hand (S→U), though less commonly seen, some studies shared a goal of learning structures in data primarily with unsupervised methods. When there were labeled data available, supervision could be exploited to improve its transferability on certain specific tasks. For instance, et al. [36] improved an unsupervised SER autoencoder by turning it semi-supervised with an adjoined supervised learning objective. These works all showed success such that combining supervised and unsupervised methods could boost performance than using either alone. Our primary goal was closer to the U→S case where we wanted to predict wellbeing labels from raw sensor data. Unfortunately, we had a major challenge that the raw data were intolerably high in dimensions. Consequently, we started with an unsupervised manner to firstly reduce the input dimensionality. That is to say, the initial state of our approach could be analogized to the S→U case. Inspired by success in both directions, we decided to insert supervision into the unsupervised framework. Here, supervision worked like a regularization or feature selection mechanism; with supervision, we were able to specify that we wanted to learn features that were related more closely with future mood, health, and stress.

We observed that with different prediction models, generalized or personalized, fed on 1-day or 7-day features, supervision never degraded the final prediction performance. When predicting stress from 7-day auto features, semi-supervised features showed a significant improvement compared to the unsupervised features. In most other prediction models, however, the advantage of supervision was not as obvious. In fact, Zhu et al. [133] pointed out that although semi-supervised learning could be
a powerful approach for many tasks, its success critically replies on certain assumptions (e.g. smoothness, cluster, and manifold assumptions). If these assumptions are violated, then semi-supervised learning may not bring as much improvement as one would normally expect. Since our data were collected in the wild, on average 6.1% of the data were missing everyday. Moreover, around 25% of the SC data on daily average suffered from low quality or noise. The data quality might be an obstacle for the semi-supervised representation learning to work well.

6.4 Understanding Auto Features and Deep 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 5.9 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. [80] 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 [130] focused on predicting happiness from multi-modal 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 [80] and [130] 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 [20]. 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.

6.5 Attention is Not All We Need

Visual attention has been widely applied to provide interpretation for deep decision-making models, such as machine translation, image captioning [?], vision-based steering angle prediction for self-driving cars [71], etc. Visual attention based interpret-
tion is able to provide interesting insights not achievable by statistical interpretation techniques. Our CVA model was based on the semi-supervised framework where it can automatically learn to fix its gaze on salient time-steps while predicting wellbeing scores. Based on electrocardiogram (ECG) data yet similarly in principle, Mousavi et al. [134] developed the ECGNET which utilized attention to find salient ECG input segments that helped detecting atrial fibrillation. Although the attention mechanism has exhibited strong performance on several empirical studies, the results from this thesis also suggest that it might require more than leveraging attention layers to form a decent understanding of a deep representation learning model based on sensor data.

First, static embedding is critical to the success of the visual attention guided interpretation. With the original representation learning model, we lose almost all spatial information of the compressed features through the stacked layers of local MLP and LSTM. Also, the bottleneck configuration was too coarse with only 8 time-steps each containing 3-hour data. In order for more accurate upsampling of the attention map, we switched from linear layers to convolutional layers. The major data compression was done by two index-preserving max pooling layers. The index information could be re-used for upsampling attention maps. The idea was originally proposed by Long et al. [135] for the fully convolutional neural networks (FCN). In another area of human healthcare, Kaul et al. [136] developed an attention-based FCN for medical image segmentation, where they also adopted the index-preserving unpooling for their upsampling stage.

Second, what to do with the upsampled attention maps remains an open question. Indeed, we were able to detect high-attention moments and extract the corresponding input segments. But what is next? The extracted data segments were merely vectors of length eight (because we have upsampled attention to one-second resolution, and
the original data were in 8 Hz). Different from the latest vision or language based applications where one can directly read off the input data \cite{137,138,139}, we are still in short of knowledge about the sensor data segments at high-attention moments. In a proceeding using attention-based autoencoder for seismic interpretation \cite{140}, the authors were faced with a similar challenge and only reported the learned autoencoder features without comprehensive interpretation. To this end, future studies should be focused on establishing causal relations within the deep models. Furthermore, it is commonly believed that cross-disciplinary collaboration should be helpful for deepening the understanding of the automatically learned features \cite{141}.

6.6 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 the temporal representation learning model 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 \cite{142}. For example, See et al. \cite{143} proposed magnitude-based weight pruning, reduced the LSTM translation model’s size by 90%, yet still kept the per-
formance untouched with re-training. Whereas, Wang et al. [144] 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. [145] 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 [146]. 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 [147]. 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 [148].

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 [149]. Second, all data have been de-identified and stored on our lab server, under
information protection measures and policies of the university. Although the temporal representation learning model 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. Other to-be-considered options to preserve privacy on distributed systems include decentralized computing, differential privacy, federated computing, etc.

6.7 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, we might be able to learn more informative features. Although in this thesis, we extracted features from three channels of sensor data separately, we will consider combining the multi-modal 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 per-
sonalized 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 [155], 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 thesis. 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 an 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.
Chapter 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.

On a passively collected and self-annotated dataset (skin conductance, skin temperature, and accelerometer data; mood, health, and stress scores), we demonstrated that under both unsupervised and semi-supervised settings, the proposed representation learning frameworks can extract better features than feature crafting. In particular, the temporal autoencoder outperformed the static autoencoder. The results obtained in our experiments show that the temporal auto features can be reconstructed to the original raw data dimensions with lower reconstruction loss than the static auto features; both the generalized LSTM and the personalized MTL models trained on the temporal features can predict mood, health, and stress scores with much smaller errors than on the static features.

We also provided evidence on the important role that personalization played in predicting subjective wellbeing status. 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.

By studying the multi-modal fusion of wearable sensor data in the context of unsupervised learning temporal features, we reached a preliminary conclusion that the trimodal early fusion can best exploit the predictive efficacy of those auto features.

Another contribution of this thesis is the attempt of interpreting deep features and prediction behaviors using correlation analysis, weight visualization, and the visual attention mechanism. Finally, we discussed computation, privacy, and ethics concerns. Further studies are recommended to enhance the interpretability of the proposed framework and realize the promising potentials of implementing this framework as a real-time ubiquitous system.
Appendix A

Full List of Crafted Baseline Features

SC features (1-25)

1. sumAUC – sum of area under the SC curve
2. sumAUCFull – sum of full area under the SC curve,
3. medianRiseTime – median rise time of SC,
4. medianAmplitude – median value of SC amplitude,
5. countPeaks – number of SC peaks,
6. sdPeaks30min – standard deviation of peak numbers within 30-minute time windows,
7. medPeaks30min – median number of peaks within 30-minute time windows,
8. percentMedPeak – percent of medium peaks,
9. percentHighPeak – percent of high peaks,
10. sclPercentOff – skin conductance level (SCL) percent off,
11. sclMaxUnnorm – unnormalized SCL maximum,
12. sclMedUnnorm – unnormalized SCL median,
13. sclMeanUnnorm – unnormalized SCL mean,
14. sclMedianNorm – normalized SCL median,
15. sclSDnorm – normalized SCL standard deviation,
16. sclMeanDeriv – normalized SCL mean derivative,
17. sumStillnessWeightedAUC – sum of SC area under the curve (AUC) weighted
by stillness,

18. sumStepsWeightedAUC – sum of SC AUC weighted by steps,
19. sumStillnessWeightedPeaks – sum of SC peaks weighted by stillness,
20. maxStillnessWeightedPeaks – maximum of SC peaks weighted by stillness,
21. sumStepsWeightedPeaks – sum of SC peaks weighted by steps,
22. medStepsWeightedPeaks – median of SC peaks weighted by steps,
23. sumTempWeightedAUC – sum of SC AUC weighted by temperature,
24. sumTempWeightedPeaks – sum of SC peaks weighted by temperature,
25. maxTempWeightedPeaks – maximum of SC peaks weighted by temperature,

**ST features (26-31)**

1. maxRawTemp – maximum of raw ST,
2. minRawTemp – minimum of raw ST,
3. sdRawTemp – standard deviation of raw ST,
4. medRawTemp – median of raw ST,
5. sdStillnessTemp – standard deviation of ST in stillness,
6. medStillnessTemp – median of ST in stillness,

**AC features (32-34)**

1. stepCount – count of steps,
2. meanMovementStepTime – mean movement step time,
3. stillnessPercent – percent of no-movement time.
Bibliography


[76] F. Gasparetti, “Personalization and context-awareness in social local search:


