Early versus Late Modality Fusion of Deep Wearable Sensor Features for Personalized Prediction of Tomorrow's Mood, Health, and Stress*

Boning Li and Akane Sano

Abstract— **Predicting mood, health, and stress can sound an early alarm against mental illness. Multi-modal data from wearable sensors provide rigorous and rich insights into one's internal states. Recently, deep learning-based features on continuous high-resolution sensor data have outperformed statistical features in several ubiquitous and affective computing applications including sleep detection and depression diagnosis. Motivated by this, we investigate multi-modal data fusion strategies featuring deep representation learning of skin conductance, skin temperature, and acceleration data to predict self-reported mood, health, and stress scores (0 – 100) of college students (N = 239). Our cross-validated results from the early fusion framework exhibit a significantly higher ($p < 0.05$) prediction precision over the late fusion for unseen users. Therefore, our findings call attention to the benefits of fusing physiological data modalities at a low level and corroborate the predictive efficacy of the deeply learned features.**

Clinical relevance— **This establishes that with automatically extracted features from multiple sensor modalities, choosing the proper scheme of fusion can reduce the errors of predicting new users' future wellbeing by as much as 13.2%.**

I. INTRODUCTION

Subjective wellbeing, the feeling of life experience being positive and satisfied (e.g. happy, healthy, calm, etc.), correlates with physiological and psychological functioning [1, 2]. Therefore, its accurate prediction grounds the successful development of early intervention systems for wellbeing enhancement. Meanwhile, ubiquitous technologies have assisted users and researchers to collect continuous biobehavioral signals unobtrusively. These signals and deep multi-modal learning enabled high performance in wellbeing prediction [3, 4].

Multi-modal learning aims to improve learning performance by leveraging different sources of data to deliver complementary and comprehensive information. Intrinsically, the combination of multiple modalities can result in richer information than unimodal data. Depending on the conducting point, multi-modal fusion can be categorized into late fusion and early fusion. While late fusion usually concatenates features that are independently extracted from each modality, early fusion extracts joint features directly from the merged raw or preprocessed data [5]. Both have demonstrated success across multi-modal learning problems including multi-scale image and speech emotion recognition [6, 7] and multi-modal depression detection [8]. On the other hand, they are facing certain challenges such as how to combine heterogeneous data with different levels of noise [9].

Meaningful and effective data representation is crucial to multi-modal learning problems. Recently, a shift has occurred from manually crafting features to automatically learning features from raw data. Yang et al. [10] used a mixture of predefined and learned extractors to compute features of multi-modal text, video, and audio data for depression level measurement. Numerous studies have proved the advantage of distributing raw sensory data in artificial neural networks to learn features through back propagation for predicting activities, depressive states, emotions, etc. [11, 12]. In this paper, we highlight the fully automatic learning and extraction of sensor features using autoencoder-based representation learning.

To choose the prevailing multi-modal fusion scheme for predicting mood, health, and stress, we compare the early and late fusion schemes based on deep representation learning. We test the predictive efficacy of the fused features using personalized regression models. We also provide an entry point to interpreting the fused deep features and explaining their predictions. This paper contributes as a first attempt to study the physical interpretation of the early versus late fusion of automatically learned wearable sensor features for predicting future wellbeing in a personalized manner.

II. METHODS

A. Dataset

Wearable sensor data including skin conductance (SC), skin temperature (ST), and acceleration (AC) were collected from 255 college students in New England in the SNAPSHOT study [13] from 2013 to 2017 for approximately 30 days to 3.5 months. SC is related with physiological arousal and sympathetic nervous activity [14]; ST has been used to measure emotion and comfort [15]; AC can reveal movement and activity-related wellbeing such as energy expenditure [16]. These data were passively collected using a wrist-worn device with a unified sampling rate of 8 Hz. To remove artifacts and environmental noise, SC data were filtered using a 32nd FIR filter with a cutoff frequency at 0.4 Hz [17], and a wavelet filter was adopted to the ST data with Symlets 4 scaling and adaptive threshold [18, 19]. The magnitude of AC was taken across three axes. Evening wellbeing scores were collected via email surveys every day at 5 pm. The self-assessed scores of mood (sad-happy), health (sick-healthy), and stress (stressed-calm) were reported on a continuous 100-point slider. Gender and Big Five Personality were also collected via standardized pre-study surveys [20].

The dataset had missing data for various reasons (e.g., sensor outage, survey incompletion, etc.). Missing sensor data were imputed with each participant's correspondent

*This work was supported by the National Institute of Health (R01GM105018), the National Science Foundation (#1840167), Samsung Electronics and NEC Corporation. The study was approved by MIT's COUHES (#1209005240) and Rice University (IRB-FY2018-451).

[1]Department of Electrical and Computer Engineering, Rice University. {boning.li, akane.sano}@rice.edu

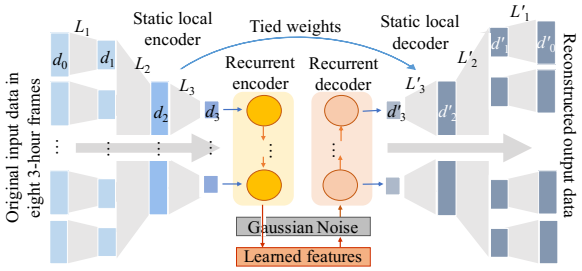


Fig. 1: Schematics of the locally connected temporal denoising autoencoder for automatic feature learning of raw sensor data, where multiple layers ($L/L'_{\{1,2,3\}}$) are stacked to propagate data through different sized hidden nodes ($d/d'_{\{1,2,3\}}$).

channel-wise mean values of the day. Days without survey responses or with more than 25% sensor data missing were discarded. After filtering and cleaning, 6391 days from 239 participants remained valid. For each participant, daily sensor data were normalized to range 0 – 1 to reduce bias.

B. Representation Learning

We composed the locally connected multilayer perceptron (LC-MLP) layers and the long short-term memory (LSTM) layers with a denoising autoencoder to form a deep representation learning framework. The loosely connected design of the first few layers preserves local information and prevents overparameterization. Temporal information in the features is enhanced by the recurrent encoding-decoding component.

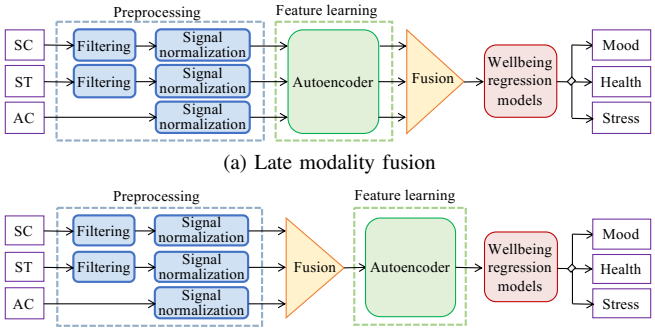
Twenty-four-hour sequences of sensor data were firstly divided into eight 3-hour input frames to address their sequential characteristics. In each input frame, we adapted the LC-MLP as the static structure. Outputs of eight frames were later joined by a many-to-one LSTM encoder, and the final hidden state of the bottleneck layer was extracted as the learned representation. With additive Gaussian noise $\sim \mathcal{N}(0, 0.1)$, this representation vector was then copied to the input of a symmetric LSTM one-to-many decoder, unrolled, and then decompressed to the input dimensions via a stack of LC-MLP symmetric to the static encoder with tied weights (Figure 1).

C. Modality Fusion

In multi-modal machine learning problems, when and how to merge modalities remained an application-dependent choice. In this paper, we compared two modality fusion schemes (Figure 2). We empirically chose 48 as the final feature dimensionality for both schemes.

1) *Late fusion*: As Figure 2a shows, 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.

2) *Early fusion*: As Figure 2b shows, 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.



(a) Late modality fusion

(b) Early modality fusion

Fig. 2: Illustration of multi-modal data fusion schemes.

D. Personalized Wellbeing Prediction

1) *Multi-task learning (MTL)*: To predict next evening's wellbeing scores from current day's sensor data, we applied the $\ell_{2,1}$-norm regularized multi-task least squares regression model [21] given by Equation 1,

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

where X_i represents the input matrix to the i-th task, and Y_i is the corresponding score. θ_i is the linear model parameters of each single task. λ_2 controls the ℓ_2-norm penalty; $\lambda_{2,1}$ controls the $\ell_{2,1}$-norm penalty. This model has been used for wellbeing prediction in [22], and it resulted in higher precision and interpretability than multi-task neural networks.

2) *Personalization in MTL*: One fundamental decision to make for MTL is the definition of *tasks*. We found that the strictly personalized models (i.e. MTL with individual participants as tasks) outperformed the non-personalized models. However, the strict MTL models have to be trained with at least one data point from any to-be-tested participants, which is not applicable to real-world solutions where new users are constantly joining. Thus, we relaxed the definition of a task to be a *group* of one or more participants. The grouping was based on their gender and personality. K-prototypes clustering [23] and Silhouette score [24] were used to create and search for the optimal number of groups. In summary, the first strategy is also called user-dependent personalization, and the second is user-independent personalization.

III. EXPERIMENTS

Our task is to predict wellbeing scores for next evening (day $t+1$, 5pm) using automatically learned features of current day's SC, ST, and AC data (day t, 0 am - day $t+1$ 0am). This study aims to investigate the effect of multi-modal fusion on the performance of representation learning and wellbeing score prediction. Cross validation was adapted for the following experiments with a split ratio of 60%, 20%, 20% respectively for train, validation, and test. The representation framework was implemented using the deep learning platform PyTorch 1.0, and the prediction models were adapted from the MALSAR toolbox [25].

A. Experiment Organization

1) *Representation learning*: Within the representation learning framework, we compared the late and early fusion approaches based on the reconstruction loss between the

original input data and the reconstructed output data. The losses were computed on the validation set at the end of each training epoch.

2) *Personalized wellbeing prediction (user-dependent)*: Using the automatically learned features, we proceeded to compare the outcomes of different modality fusion schemes with regard to the personalized wellbeing prediction. Individual participants were treated as unique tasks in MTL prediction models. We compared all combinations of three modalities (SC, ST, AC, SC+ST, SC+AC, ST+AC, SC+ST+AC).

3) *Personalized wellbeing prediction (user-independent)*: 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.

B. Evaluation Metrics

We used the Mean Squared Errors (MSE) to evaluate the representation learning performance and Mean Absolute Errors (MAE) to evaluate the performance of wellbeing prediction models. Smaller errors indicate higher performance.

C. Statistical Test

Analysis of Variance (ANOVA) [26] was used to test for significant differences among two or more groups of results, followed by a Tukey HSD test [27] to find group(s) of results that were significantly different from the others.

D. Analysis of the learned features and prediction models

To understand the sources of difference between prediction behaviors based on the late features and the early features, we endeavored to identify, analyze, and interpret the most contributing late and early features in the personalized models.

To find the critical features, we looked for features that resulted in diverse weight coefficients across different tasks. We clustered the weight vectors of 239 individual-based tasks using K-means and Silhouette score evaluation, similar to the principle described in II-D.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. For each feature, if its weight coefficients had statistically different distributions in the two clusters, and if opposite signs were observed in cluster means, it would be considered as critical to personalization, or a *critical feature*.

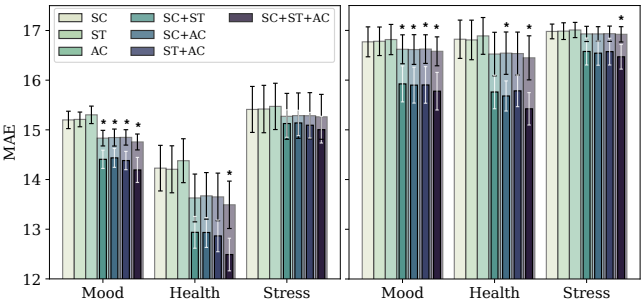
To interpret the critical features, we manually defined 34 crafted features of the raw sensor data [22]. They were computed from four non-overlapping time frames, namely 0-3H, 3-10H, 10-17H, and 17H+, resulting in 136 crafted features daily. Pearson correlation was then computed between the learned features and the crafted features. By ranking the statistically significant correlations, we could depict some intuition of the learned features and prediction behaviors.

IV. RESULTS

We compared the reconstruction MSE loss via late and early fusion of modalities during the representation learning process. 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}).

Moving forward to predicting wellbeing score, a coincident pattern was observed. The early fusion features always resulted in statistically equivalent or significantly lower errors compared to corresponding late fusion features, as Figure 3 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.



(a) User dependent (b) User independent

Fig. 3: 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.

We conducted the feature and prediction analysis to our best performing trimodal health model (user-dependent), as described in Section III-D. 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.

By summarizing the top correlations between critical auto features and hand-crafted features in Table I, we found that the critical early features had stronger correlations with ST and SC medians around evening. The top three correlations with the critical late feature were found in AC stillness percentage and step counts before and after mid-night, where step count 0-3H could be related to sleep and step count 17H+ could indicate physical activities.

V. DISCUSSION

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

TABLE I: Top-3 highest correlations between critical fusion features and hand-crafted features. All displayed correlation coefficients had p-value $< 7.3\times10^{-5}$ as the adjusted significance threshold. (* E.F. = critical early fusion feature; L.F. = critical late fusion feature.)

E.F.*-1	E.F.-2	E.F.-3	E.F.-4	L.F.*-1
0.31 17H+: ST median (stillness)	0.35 17H+: ST median	0.22 17H+: SC z-score median	0.31 17H+: ST median (stillness)	0.11 17H+: AC stillness percent
0.31 17H+: ST median	0.35 17H+: ST median (stillness)	0.21 17H+: ST median (stillness)	0.30 17H+: ST median	-0.10 0-3H: AC step count
0.16 17H+: SC z-score median	-0.16 10-17H: ST minimum	0.20 17H+: ST median	-0.16 10-17H: ST minimum	-0.10 17H+: AC step count

be smoother and more coherent than with the late fusion. According to [28] 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. [29, 30]) 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 [31] and pain [32]. Prior work has also revealed strong links from social interaction to happiness [33], suggesting that exploiting phone and other ubiquitous data is very likely to boost our model's performance.

This study is open to further improvement. Our results were solely based on a college student dataset. Re-validation on other datasets is needed, including healthy and in-treatment populations. We will study deep feature learning from other types of data such as phone usage, location, and weather logs. Our explanation of the learned features was also limited, because autoencoders could capture other predictive information beyond our current vocabulary.

VI. CONCLUSIONS

We studied the multi-modal fusion of wearable sensor data in the context of learning features that can forecast college students' subjective mood, health, and stress. Early fusion can promote the unsupervised learning of sensor features. Without prior knowledge of any target students, the lowest prediction MAE was reached at 15.4 (S.D.=0.3) in predicting their health using the early fused SC+ST+AC features. There was a 6.1% improvement compared to the late fused SC+ST+AC features and a 13.2% improvement to the unimodal features.

REFERENCES

[1] C. L. Keyes and J. L. Magyar-Moe. "The measurement and utility of adult subjective well-being." In: (2003).

[2] J.-E. De Neve et al. "The objective benefits of subjective well-being". In: *World happiness report* (2013).

[3] N. Jaques et al. "Multi-task, multi-kernel learning for estimating individual wellbeing". In: *Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec*. Vol. 898. 2015.

[4] A. Tsakalidis et al. "Combining heterogeneous user generated data to sense well-being." In: The COLING 2016, 2016.

[5] T Baltrusaitis et al. "Multimodal machine learning". In: *The Handbook of Multimodal-Multisensor Interfaces* 2 (2018).

[6] S. Zhao et al. "Approximating discrete probability distribution of image emotions by multi-modal features fusion". In: *Transfer* 1000.1 (2017), pp. 4669–4675.

[7] J. Sebastian and P. Pierucci. "Fusion techniques for utterance-level emotion recognition combining speech and transcripts". In: *Proc. Interspeech*. 2019, pp. 51–55.

[8] A. Samareh et al. "Predicting depression severity by multi-modal feature engineering and fusion". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[9] T. Baltrušaitis et al. *Multimodal Machine Learning: A Survey and Taxonomy*. 2017. arXiv: 1705.09406 [cs.LG].

[10] L. Yang et al. "Multimodal measurement of depression using deep learning models". In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 2017, pp. 53–59.

[11] R. Hecht-Nielsen. "Theory of the backpropagation neural network". In: *Neural networks for perception*. Elsevier, 1992, pp. 65–93.

[12] F. Ordóñez and D. Roggen. "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition". In: *Sensors* 16.1 (2016), p. 115.

[13] A. Sano et al. "Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: Observational study". In: *Journal of medical Internet research* 20.6 (2018), e210.

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

[15] D. Wang et al. "Observations of upper-extremity skin temperature and corresponding overall-body thermal sensations and comfort". In: *Building and Environment* 42.12 (2007), pp. 3933–3943.

[16] A. K. Chowdhury et al. "Deep learning for energy expenditure prediction in pre-school children". In: (2018).

[17] A. Sano et al. "Quantitative analysis of wrist electrodermal activity during sleep". In: *International Journal of Psychophysiology* 94.3 (2014), pp. 382–389.

[18] K Palanisamy et al. "Multiple physiological signal-based human stress identification using non-linear classifiers". In: *Elektronika ir elektrotechnika* 19.7 (2013), pp. 80–85.

[19] X.-P. Zhang and M. D. Desai. "Adaptive denoising based on SURE risk". In: *IEEE signal processing letters* 5.10 (1998), pp. 265–267.

[20] O. P. John and S. Srivastava. "The Big-five Trait Taxonomy: History, Measurement, and Theoretical Perspectives". In: 1999.

[21] R. Caruana. "Multitask Learning". In: *Machine Learning* 28.1 (1997), pp. 41–75. ISSN: 1573-0565. DOI: 10.1023/A:1007379606734.

[22] H. Yu et al. "Personalized Wellbeing Prediction using Behavioral, Physiological and Weather Data". In: *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI) (IEEE BHI 2019)*. Chicago, USA, May 2019.

[23] Z. Huang. "Extensions to the k-means algorithm for clustering large data sets with categorical values". In: *Data mining and knowledge discovery* 2.3 (1998), pp. 283–304.

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

[25] J. Zhou et al. "MALSAR: Multi-task learning via structural regularization". In: *Arizona State University* (2011).

[26] G. R. Iversen et al. *Analysis of variance*. 1. Sage, 1987.

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

[28] H. Gunes and M. Piccardi. "Affect recognition from face and body: early fusion vs. late fusion". In: *2005 IEEE international conference on systems, man and cybernetics*. Vol. 4. IEEE. 2005, pp. 3437–3443.

[29] J. L. Stoops. "A possible connection between thermal comfort and health". In: (2004).

[30] S. Y. Sim et al. "Estimation of thermal sensation based on wrist skin temperatures". In: *Sensors* 16.4 (2016), p. 420.

[31] H. Pazderka-Robinson et al. "Electrodermal dissociation of chronic fatigue and depression: evidence for distinct physiological mechanisms". In: *International Journal of psychophysiology* 53.3 (2004), pp. 171–182.

[32] E. Syrjälä et al. "Skin Conductance Response to Gradual-Increasing Experimental Pain". In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2019, pp. 3482–3485.

[33] N. Jaques et al. "Predicting students' happiness from physiology, phone, mobility, and behavioral data". In: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2015, pp. 222–228.