More to Less (M2L): Enhanced Health Recognition in the Wild with Reduced Modality of Wearable Sensors

Huiyuan Yang[1], Han Yu[1], Kusha Sridhar[1], Thomas Vaessen[2], Inez Myin-Germeys[2] and Akane Sano[1]

***Abstract*— Accurately recognizing health-related conditions from wearable data is crucial for improved healthcare outcomes. To improve the recognition accuracy, various approaches have focused on how to effectively fuse information from multiple sensors. Fusing multiple sensors is a common choice in many applications, but may not always be feasible in real-world scenarios. For example, although combining biosignals from multiple sensors (i.e., a chest pad sensor and a wrist wearable sensor) has been proved effective for improved performance, wearing multiple devices might be impractical in the free-living context. To solve the challenges, we propose an effective more to less (M2L) learning framework to improve testing performance with reduced sensors through leveraging the complementary information of multiple modalities during training. More specifically, different sensors may carry different but complementary information, and our model is designed to enforce collaborations among different modalities, where positive knowledge transfer is encouraged and negative knowledge transfer is suppressed, so that better representation is learned for individual modalities. Our experimental results show that our framework achieves comparable performance when compared with the full modalities. Our code and results will be available at https://github.com/comp-well-org/More2Less.git.**

I. Introduction

Wearable sensors are unobtrusive, affordable and user-friendly, making them suitable for continuous and ubiquitous monitoring of individual's physiological and behavioral profiles in the free-living context and providing valuable insights into individuals' health and fitness status for health and medical applications. These advantages have attracted more and more researchers to adopt multimodal machine learning to various wearable devices for better health monitoring and interventions, as fusing data from different modalities can aggregate more information, therefore outperforming their unimodal counterparts. Huang et al. [4] provide appealing formal guarantees about the performance advantages of multimodal learning in comparison with unimodal learning using theoretical proofs. Some examples of multimodal learning frameworks proposed in the literature have fused audio and visual information for speech recognition [13], improved word embeddings with both text and visual information [8] and learned joint representations from text, visual and audio modality for sentiment analysis. Moving our attention to the field of wearable sensors, researchers have attempted to infer health related constructs or clinical events from multimodal signals (*i.e., fitness trackers, smartwatches and smartphones*), such as mental health and wellbeing monitoring [18, 2, 5], seizure forecasting[10], COVID-19 detection [11], etc. Other work like [15] exploits the relationship among different modalities for better representation learning from wearable data, [7] uses additional sensors to improve single-sensor based complex activity recognition.

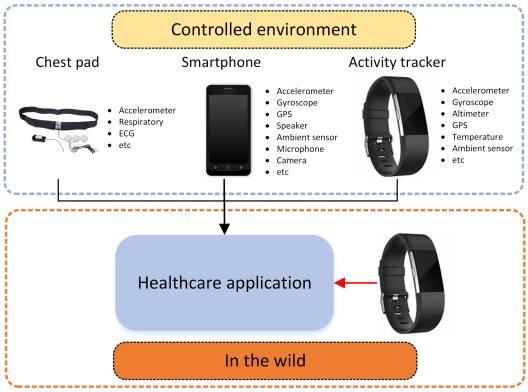

Fig. 1. Illustration of the targeted challenge. In the controlled setting, we are usually able to collect various signals using multiple devices, such as standalone sensors, phone and fitbit, but less number of sensors or devices is more convenient for users in real-world settings. *How can we design a framework that can leverage the benefits of multimodal data during training but maintain the robustness of model performance with reduced number of modalities during testing?*

Although research using multimodal modeling and wearable sensor technologies has progressed with the advent of deep learning, only a small number of these studies have been successfully applied in our society. Some of the challenges that hinder the widespread adoption of wearable devices in healthcare applications include data acquisition and preprocessing, feature extraction, and model selection. However, one specific challenge that we face in many real-world applications is the reduced availability of sensor modalities or devices in deployment compared to model training, and the challenge is understudied in the literature.

A common assumption for most works is that we have access to an equal number of sensors in both training and testing. However, such an assumption does not hold true in many circumstances in real-world scenarios. For example, as shown in Fig.1, it is usually feasible to collect multiple modalities of data from study participants using different sensors (i.e., standalone sensors, chest sensors, wearables) in the controlled environment such as laboratory experiments. Therefore, we may be able to develop a robust multimodal model through effective fusion of complementary informa-

*This work was supported by NSF #1840167 and #2047296

[1] H. Yang, H. Yu, K. Sridhar and A. Sano are with the Department of Electrical Computer Engineering, Rice University, Houston TX 77005, USA. Email: {hy48, hy29, kh82, Akane.Sano}@rice.edu

[2]T. Vaessen and I. Myin-Germeys are with the Department of Neurosciences, KU Leuven, Leuven, Belgium. Email: {thomas.vaessen, inez.germeys}@kuleuven.be

tion from multiple modalities. However, in real deployment, using less number of sensors or devices is preferred, as we can therefore minimize user burden, energy consumption, or device size/cost. (*i.e., fitbit only*).

Therefore, it is critical to bridge the gap between the models developed using multiple sensors during development and the models using less number of sensors during deployment in the wild. In this work, we present an effective framework, which not only leverages the complementary information of multiple modalities during training, but has the ability to provide inference with fewer modalities during testing (*simulation of the model real-world deployment*). More specifically, an adaptive gate is designed for the multi-modalities, which will control the direction and intensity of knowledge transfer among modalities. Therefore, positive knowledge transfer is encouraged, while negative knowledge transfer is suppressed. After training, we can thus expect improved performance for individual modality. Our main contributions are:

- We propose an effective M2L framework that not only can leverage the complementary information of multiple modalities during training but also provide inference with fewer modalities during testing.
- We conduct extensive experiments using two wearable datasets, and demonstrate that our framework can benefit from the multimodal training, achieving comparable performance in testing with reduced modalities.

II. Propose Method

We propose an effective **M**ore to **L**ess (**M2L**) framework, which is designed to learn robust representations for each modality through leveraging the specific strengths existed in different modalities. This is accomplished by utilizing a cooperative learning strategy, where a weak network learns representations from a stronger network through knowledge distillation. More specifically, assuming M modalities and M classifier networks with similar architectures are available during training, with each classifier trained with its own data, it will also try to learn representations from other classifiers that show better performance than itself. The knowledge sharing is applied to the representations through minimizing the distance between the two features in the embedding space. In order to guarantee positive knowledge transferring, an adaptive regularizer is applied to ensure that knowledge only transfers from more accurate modality networks to those with less accuracy, and not the other way.

A. Classification Loss and Feature Distance

Let $\mathcal{D} = \{(x_i, y_i)\}^N$ be a multimodal dataset having N training samples. Each sample x_i represents the data of M available modalities, $x_i = \{x_i^1, x_i^2, \dots x_i^M\}$, and y_i represents its label. For each modality of x_i^m, a backbone network $\mathbf{f}^m(\cdot)$ is used to map the input into feature space: $\mathbf{f}^m : x_i^m \rightarrow F_i^m$, and $F_i^m \in \mathbb{R}^d$. The supervised classification loss with respect to a specific modality and network

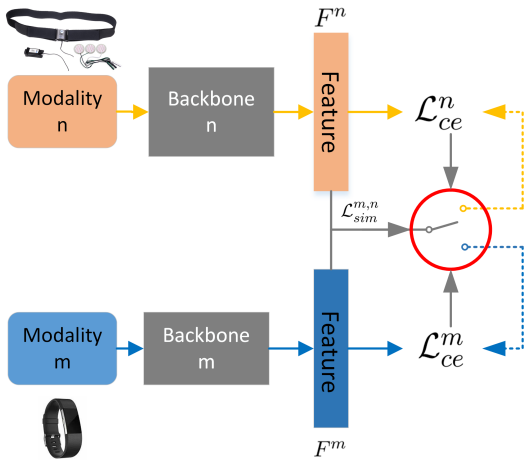

Fig. 2. Framework of the proposed M2L method. Assuming we want to train backbone m for the mth modality with the information shared in nth modality. First, the cross-entropy classification loss, $\mathcal{L}_{ce}^m$ and $\mathcal{L}_{ce}^n$ are calculated for each modality. Next, $\rho^{n\rightarrow m}$ and $\rho^{m\rightarrow n}$, determining the direction and intensity for distillation-based regularization, is calculated through comparing the two loss functions $\mathcal{L}_{ce}^m$ and $\mathcal{L}_{ce}^n$. The full objective function for training backbone m for the mth modality is the weighted combination of $\mathcal{L}_{ce}^m$ and $\mathcal{L}_{sim}^{m,n}$. After training, the proposed framework can be run with reduced modalities, while leveraging the benefits of multimodal training.

(i.e. mth modality) is defined as:

$$\mathcal{L}_{ce}^m = -\frac{1}{N}\sum_{i=1}^{N}\Big\{y_i \times \log(\hat{y}_i) + (1-y_i)\times\log(1-\hat{y}_i)\Big\} \quad (1)$$

To benefit from the complementary information in the multiple modalities, we encourage the networks of the mth and the nth to share their own advantages with each other. This can be done through minimizing their distance in the feature space, but experiments suggest that directly minimizing the L2 distance in the feature space will lead to unstable training. As a result, we choose the cosine similarity as the metric:

$$\mathcal{L}_{sim}^{m,n} = \frac{1}{N}\sum_{i=1}^{N}\frac{F_i^m \cdot (F_i^n)^T}{||F_i^m|| \cdot ||F_i^n||} \quad m,n \in \{1,2,\dots M\} \quad (2)$$

B. Transferring Positive Knowledge

As mentioned before, different modalities may convey varied information, and some modalities may provide weak features as compared to the others and vice versa. In addition, even the strong modalities may sometimes have corrupted samples such as noise in the training dataset. With these cases in mind, it is desirable to develop a method that encourages positive knowledge transfer between the networks while avoiding negative transfer. Such a mechanism is implemented as $\rho(\cdot)$ in our framework. For example, as shown in Fig.2, $\rho^{n\rightarrow m}$ regulates the direction and intensity of transferring knowledge from modality n to modality m.

Assume $\mathcal{L}_{ce}^m$ is the classification losses of the networks m. Next, let $\Delta\mathcal{L}^{i\rightarrow m} = \mathcal{L}_{ce}^m - \mathcal{L}_{ce}^i, i \in \{1,2,\dots M\}$ be their difference. A positive $\Delta\mathcal{L}^{i\rightarrow m}$ indicates that network i works better than network m. Hence, in the training of network m, we want $\rho^{i\rightarrow m}$ to open the gate and transfer knowledge from

network i to network m, where the strength is conditioned on the value of $\Delta\mathcal{L}^{i\to m}$. On the other hand, a negative $\Delta\mathcal{L}^{i\to m}$ indicates that network i is weaker than network m, so we want to avoid the knowledge transfer by setting $\rho^{i\to m}$ as 0. The regularizer for training modality m with the assistance of other modalities is defined as below:

$$\rho^{i\to m} = \begin{cases} e^{\beta\Delta\mathcal{L}^{i\to m}} - 1 & \Delta\mathcal{L}^{i\to m} > 0 \\ 0 & \Delta\mathcal{L}^{i\to m} \leq 0 \end{cases} \quad i \in \{1, 2, \dots M\} \tag{3}$$

where β is a positive hyper-parameter, which is used to control the strength of knowledge transferring.

C. Full Objective Function

Combining all the loss functions together, our full objective function for the training of network m corresponding to the mth modality in a dataset containing M modalities is defined as follows:

$$\mathcal{L}^m_{all} = \mathcal{L}^m_{ce} + \lambda \cdot \sum_{n=1, n\neq m}^{M} \rho^{n\to m} \cdot \left(1 - \mathcal{L}^{m,n}_{sim}\right) \tag{4}$$

where λ is a positive regularization parameter. Fig.2 shows an overview of how the features of nth modality assist the learning procedure of the mth modality. After training, the cross-modality knowledge transferring module is not needed any more, therefore the individual modality can be run independently.

III. Experiments

We use two wearable multimodal datasets to evaluate our proposed framework.

A. Data

SMILE[14]: wearable sensor and self-report data collected from 41 healthy participants (36 females and 5 males) in a 10-day study. Two types of wearable sensors were used to collect galvanic skin response (GSR) (a *wrist-worn* device, Chillband, IMEC, Belgium, sampling rate: 256 Hz) and electrocardiogram (ECG) (chest patch sensor, Health Patch, IMEC, Belgium, 256 Hz). Both time and frequency domain statistical features related to human stress status[6, 12] were extracted every minute from GSR (12 features) and ECG data (8 features) (see more about features in [14, 18]). Self-reported stress levels (0 ("not at all") to 6 ("very")) were also collected 10 times daily as ecological momentary assessment that were spaced out roughly 90 minutes apart. We set stress levels greater than 1 as positive examples (55%) and others as negative examples (45%) in our experiments. We used 1 hour of GSR and ECG data (1 minute $\times$ 60 steps) to infer upcoming stress labels.

TILES[9]: wearable, smartphone, and survey data collected from over 200 hospital workers (31.1% of the participants were male and 68.9% were female, and age: 21 - 65 years old). We use heart rate and step count data collected with the Fitbit Charge 2 (sampled every 1 minute) and ECG data collected with the OMSignal smart garment (15-second long ECG signal in 250 Hz every 5 minutes). We extracted 25 time and frequency-domain ECG features (see more in [1]) and resampled Fitbit data every 5 minutes to align with the ECG features. Self-reported stress levels were annotated by participants in a 5-point scale, which is further binarized via a similar procedure as used in [3] using the average z-score of individual's stress levels. We used 2 hours of the data (5 minute $\times$ 24 steps) to infer upcoming stress labels.

For both datasets, we randomly split 70% of the participants of the data as a training set, and the rest as a test set to conduct subject-independent experiments, where data collected from individual subjects can only appear in either training or testing, but not both.

B. Implementation Details

To model the long sequential data, we use long short-term memory (LSTM) as backbone. Each of the LSTM models is a two layers of LSTM with 64 as the number of features in the hidden state. Dropout rate is set as 0.5 during training, and 0 during testing. For the hypterparameters, λ is set to 0.05, and $\beta = 2$.

The networks are trained from scratch for all the experiments, using Adam optimizer with an initial learning rate of 0.001. The learning rate is decayed after every 10 epochs by 0.1. Batch size is set 100, and the model is trained for 50 epochs with early stopping. We start with pretraining individual modalities for 20 epochs, and then continue training with the knowledge transfer loss. We implement the model with the Pytorch framework and perform training and testing on the NVIDIA GeForce 3090 GPU.

C. Results

Various experiments are conducted to evaluate the performance of the proposed method. Both accuracy and F1-score are reported in our experiments for performance evaluation. First, we verify our claim that different modalities convey varied information. As shown in Table I, the accuracy changes over different modalities *i.e., GSR, ECG and fitbit*. We also calculate the consistency ratio of individual modality based prediction, which is calculated by dividing the total number of testing examples by consistent prediction of two individual modality. The consistency ratio is only around 60% in both datasets, indicating varied information among different modalities.

TABLE I
ACCURACY AND CONSISTENCY RATIO IN TWO DATASETS.

Dataset	Modality	Acc	Ratio
SMILE	GSR	53.9	59.0
	ECG	50.8	
TILES	ECG	55.2	59.2
	fitbit	57.5	

The performance evaluation on the SMILE dataset is reported in Table II. GSR-based model achieves 3% higher accuracy than ECG-based model, while the early fusion of GSR+ECG does not always perform better than the individual modality, indicating that even though the multiple modalities contain rich and complementary information, the benefits will not be achieved without a carefully designed fusion mechanism. Compared to the models trained and

TABLE II

F1 SCORES AND ACCURACY OF THE PROPOSED METHOD ARE REPORTED ON THE SMILE DATASET FOR STRESS DETECTION IN THE WILD.

Method	Training modality	Testing modality	Reduced modality	Acc	F1
LSTM	GSR	same	✗	53.9	64.1
	ECG	same	✗	50.8	61.5
LSTM (fusion)	GSR+ECG	same	✗	52.6	58.8
M2L	GSR + ECG	GSR	✓	56.1	66.8
	GSR + ECG	ECG	✓	53.1	64.5

TABLE III

F1 SCORES AND ACCURACY OF THE PROPOSED METHOD WITH THE TILES DATASET FOR STRESS DETECTION IN THE WILD.

Method	Training modality	Testing modality	Reduced modality	Acc	F1
LSTM	ECG	same	✗	55.7	65.8
	fitbit	same	✗	57.4	66.7
LSTM (fusion)	ECG+fitbit	same	✗	54.6	61.5
Paper [16]	BR	same	✗	59.5	56.9
Paper [17]	HRV	same	✗	60.6	58.2
M2L	ECG+fitbit	ECG	✓	57.5	72.2
	ECG+fitbit	fitbit	✓	58.6	70.3

tested with a single modality (*GSR or ECG*), our model can be trained with multiple modalities (*GSR and ECG*) and tested with reduced modality (*GSR only or ECG only*), showing significant improvement (2.2% and 2.3% higher in accuracy respectively). The improved performance for reduced modality testing demonstrates the effectiveness of our proposed method.

Experiment results using the TILES dataset are shown in Table III, and we observe 1.8% and 1.2% improvement in accuracy when compare our model with LSTM on the ECG and fitbit modality respectively. Comparing with recent stress detection works [16, 17], which were trained and tested in TILES datasets using breathing rate (BR) and heart rate variability (HRV) respectively collected with OMSignal Garment (details about training and testing and label settings are not described so we are not able to reproduce their experiments), our model not only achieves comparable performance, but also is able to run with reduced modality (*an inexpensive solution, i.e., fitbit*), while all other methods need exactly the same modalities between training and testing.

IV. CONCLUSION

The reduced number of wearable sensors/devices during deployment/testing in the wild environment compared to model training is a challenge that hinders the adoption of wearable devices for healthcare. To solve this problem, we present an effective M2L framework that can not only leverage the complementary information of multiple modalities during training, but also provide inference with reduced modalities during testing, while achieving comparable performance when compared with full modalities, therefore, bridging the gap between model development and model deployment in the wild. It is worth noting that the proposed framework works for three or more modalities as well.

Although the improved performance is shown, our proposed method is still limited by the use of hand-crafted features and the ability to generalize to new subjects. In the future, we plan to investigate both deep learning based feature learning and unsupervised personalization, where a general model can be automatically adapted to individual through utilizing a small amount of unlabeled data.

REFERENCES

[1] Robin Champseix. *Heart Rate Variability Analysis*. `https://github.com/Aura-healthcare/hrv-analysis`. 2018.

[2] Mohamed Elgendi and Carlo Menon. "Assessing anxiety disorders using wearable devices: Challenges and future directions". In: *Brain sciences* 9.3 (2019), p. 50.

[3] Amr Gaballah et al. "Context-Aware Speech Stress Detection in Hospital Workers Using Bi-LSTM Classifiers". In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 8348–8352.

[4] Yu Huang et al. "What Makes Multimodal Learning Better than Single (Provably)". In: *arXiv preprint arXiv:2106.04538* (2021).

[5] Natasha 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, p. 3.

[6] Hye-Geum Kim et al. "Stress and heart rate variability: a meta-analysis and review of the literature". In: *Psychiatry investigation* 15.3 (2018), p. 235.

[7] Paula Lago et al. "Using additional training sensors to improve single-sensor complex activity recognition". In: *Int. Symposium on Wearable Computers*. 2021, pp. 18–22.

[8] Junhua Mao et al. "Training and evaluating multimodal word embeddings with large-scale web annotated images". In: *arXiv preprint arXiv:1611.08321* (2016).

[9] Karel Mundnich et al. "TILES-2018, a longitudinal physiologic and behavioral data set of hospital workers". In: *Scientific Data* 7.1 (2020), pp. 1–26.

[10] Mona Nasseri et al. "Ambulatory seizure forecasting with a wrist-worn device using long-short term memory deep learning". In: *Scientific reports* 11.1 (2021), pp. 1–9.

[11] Giorgio Quer et al. "Wearable sensor data and self-reported symptoms for COVID-19 detection". In: *Nature Medicine* 27.1 (2021), pp. 73–77.

[12] Nandita Sharma and Tom Gedeon. "Objective measures, sensors and computational techniques for stress recognition and classification: A survey". In: *Computer methods and programs in biomedicine* 108.3 (2012), pp. 1287–1301.

[13] Bowen Shi et al. "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction". In: *arXiv preprint arXiv:2201.02184* (2022).

[14] Elena Smets. "Towards large-scale physiological stress detection in an ambulant environment". In: (2018).

[15] Dimitris Spathis et al. "Self-supervised transfer learning of physiological representations from free-living wearable data". In: *Proceedings of the Conference on Health, Inference, and Learning*. 2021, pp. 69–78.

[16] Abhishek Tiwari et al. "Breathing rate complexity features for "in-the-wild" stress and anxiety measurement". In: *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE. 2019, pp. 1–5.

[17] Abhishek Tiwari et al. "Stress and anxiety measurement" in-the-wild" using quality-aware multi-scale hrv features". In: *IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2019, pp. 7056–7059.

[18] Han Yu et al. "Modality Fusion Network and Personalized Attention in Momentary Stress Detection in the Wild". In: *9th Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2021, pp. 1–8.