Semi-Supervised Learning for Wearable-based Momentary Stress Detection in the Wild

HAN YU and AKANE SANO*, Rice University, USA

Physiological and behavioral data collected from wearable or mobile sensors have been used to estimate self-reported stress levels. Since stress annotation usually relies on self-reports during the study, a limited amount of labeled data can be an obstacle to developing accurate and generalized stress-predicting models. On the other hand, the sensors can continuously capture signals without annotations. This work investigates leveraging unlabeled wearable sensor data for stress detection in the wild. We propose a two-stage semi-supervised learning framework that leverages wearable sensor data to help with stress detection. The proposed structure consists of an auto-encoder pre-training method for learning information from unlabeled data and the consistency regularization approach to enhance the robustness of the model. Besides, we propose a novel active sampling method for selecting unlabeled samples to avoid introducing redundant information to the model. We validate these methods using two datasets with physiological signals and stress labels collected in the wild, as well as four human activity recognition (HAR) datasets to evaluate the generality of the proposed method. Our approach demonstrated competitive results for stress detection, improving stress classification performance by approximately 7% to 10% on the stress detection datasets compared to the baseline supervised learning models. Furthermore, the ablation study we conducted for the HAR tasks supported the effectiveness of our methods. Our approach showed comparable performance to state-of-the-art semi-supervised learning methods for both stress detection and HAR tasks.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**; • **Computing methodologies → Machine learning**.

Additional Key Words and Phrases: Wearable Data, Stress Detection, Semi-Supervised Learning, Time-Series Learning

ACM Reference Format:
Han Yu and Akane Sano. 2023. Semi-Supervised Learning for Wearable-based Momentary Stress Detection in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 2, Article 80 (June 2023), 23 pages. https://doi.org/10.1145/3596246

1 INTRODUCTION

Stress is common and complex. It can benefit people under certain circumstances and increase resilience. Exposure to moderate levels of stress can be beneficial as it can prepare an organism to deal with challenges [13]. On the other hand, stress has also been associated with an increased risk for many somatic and mental illnesses [1], increasing risks for cardiovascular health issues [22] and suppressing the human immune system [23]. Effectively detecting moments of stress in real life may help an individual regulate their stress behaviorally to promote resilience and wellbeing.

Widespread portable devices bring the potential to measure human emotion, including stress levels, using passively sensed data. For example, wearable sensors and smartphones have enabled real-time monitoring of physiological and behavioral data such as body acceleration, skin conductance, skin temperature, heart rate, and phone usage. Prior studies have developed machine learning models to measure momentary self-reported stress levels using physiological, behavioral sensors, and survey features [10, 16, 37, 39]. Although these prior

Authors' address: Han Yu, hy29@rice.edu; Akane Sano, akane.sano@rice.edu, Rice University, 6100 Main Street, Houston, Texas, USA, 77005.

2474-9567/2023/6-ART80 $15.00
https://doi.org/10.1145/3596246

studies provided promising results in stress estimation, we could further improve the performances by addressing challenges in data. In the studies introduced above, the number of self-reported labels is usually limited. On the other hand, wearable devices can collect millions of data samples throughout the study period. The aforementioned studies focused on using data aligned with stress labels, which resulted in information loss when discarding unlabeled data. In alleviating the sparse label issue, semi-supervised algorithms have been widely studied, especially for the computer vision applications [5, 12, 26, 32, 41, 50, 51]. Moreover, inspired by the ideas of these image-based studies, researchers developed methods, including both semi-supervised learning and self-supervised learning algorithms, of leveraging unlabeled wearable data for applications including human activities recognition [21, 36, 45, 56], cardiovascular risk prediction [3], and stress detection [38, 49].

The aforementioned studies achieved success in their respective fields, and they all illustrated that the information from unlabeled data could boost the performance for the supervised tasks. However, there are issues needed to be resolved when utilizing these methods on wearable data from stress detection because of the inconsistency of the domain challenges and knowledge. First, the labeled and unlabeled data collected in the wild can be in different data distributions, which breaks a critical prerequisite of the aforementioned methods in leveraging unlabeled data for helping the supervised tasks. For example, massive amounts of sleep time samples exist in the databases, whereas the labeled data are collected during non-sleep periods. Secondly, as an essential tool of the methods mentioned above, the data augmentation techniques can be under-studied in the wearable data domain, and different augmentations contribute differently to the robustness of the model [52]. Therefore, in this work, we proposed a two-stage semi-supervised learning framework with modules to alleviate the above-mentioned issues. We designed an active sampling method to explicitly select unlabeled samples that obey the same distribution as the labeled data. We also introduced the averaging mechanism with a consistency regularization method to alleviate the detriment of the learning parameters from improper augmentations. Our contribution can be summarized as follows:

- We developed a two-stage semi-supervised learning framework that includes an auto-encoder pre-training method and consistency regularization to leverage both labeled and unlabeled data for robust model training.
- We propose an active sampling approach for selecting unlabeled data in semi-supervised learning to reduce the distribution differences between the labeled and unlabeled data for the stress detection system.
- We evaluated the proposed methods using two datasets, including multimodal sensor data and momentary stress labels collected in the wild. Additionally, we conducted an ablation study on four human activity recognition (HAR) datasets. We observed clear improvements in model performance using the proposed methods compared to the baselines.
- We conducted empirical experiments to examine state-of-the-art (SOTA) methods from other domains in wearable-based stress detection tasks. Our proposed method showed competitive performance compared to the SOTA methods.

2 RELATED WORK

Traditionally, stress has been measured using surveys. For example, the Perceived Stress Scale (PSS) was developed for measuring perceived stress over the past month [8]. The Holmes and Rahe Stress Scale is another survey instrument that adds up the self-reported events in the prior year that could lead to stress, estimates the total amount of stress in the past year, and determines whether people are at risk of becoming sick [19]. However, a drawback of these traditional surveys is that they might be cumbersome for individuals to complete, especially when the studies last for a long time. The resulting fatigue may make survey answers less reliable.

2.1 Stress Detection Using Wearable Data

With the development of mobile phones and wearable devices, accessing users' physiological and behavioral data in daily life settings has become a boost in monitoring human mental status. Machine learning has enabled us to develop models to learn patterns from data samples and has already brought benefits to ubiquitous computing applications. Multi-modal data from wearable sensors, mobile phones, and other smart devices have been widely used with machine learning in estimating human emotion as well as momentary stress levels [4, 10, 16, 18, 24, 27, 37, 39, 54]. Yang *et al.* proposed an attention-based LSTM system to fit data from smartphones and wristbands and predicted the participants' positive or negative emotional states with an accuracy of 89% [54]. Hinkle *et al.* leveraged multi-modal physiological signals - such as Electrocardiogram (ECG) and Electroencephalogram (EEG) - to detect human emotions by classifying binary classes of arousal and valence [18]. They achieved an accuracy of 89% with an SVM model. Bari *et al.* used wearable physiological and inertial sensors to record data of 38 employees for detecting the stressful status in human conversation, they extracted and modeled features using a random forest model with an F1 score of 0.83 [10]. Li *et al.* developed a deep learning method of using an auto-encoder to extract features from raw physiological data including electrodermal activity (EDA), body movement, and skin temperature (ST). They predicted stress status in a regularization task with an absolute error of 15.0 out of 100 [27]. Shi *et al.* collected 22 subjects' ECG, galvanic skin response (GSR), respiration (RIP), and ST data using wearable sensors [39]. Each subject in the study was exposed to a protocol, including four stressors and six rest periods, and stress labels were collected before and after each stressor/rest period through interviews. The authors proposed a personalized SVM algorithm to classify binary stress labels (low/high), which provided a precision of 0.68 with a recall of 0.80. In these studies, while physiological data were collected continuously using wearable sensors, human stress labels were collected using questionnaires. To match the sensor data with sparse labels, the collected physiological data were downsampled to align momentary features with stress labels, which [37, 39] caused a loss of information.

2.2 Semi-supervised Learning

To overcome the difficulty in learning models with a small number of labels, numerous semi-supervised learning methods have been developed to leverage massive unlabeled samples [5, 12, 26, 32, 41, 50, 51]. For example, Laine *et al.* proposed an Π-model to infer the predictions from two transformed images from a single unlabeled source and regularized the discrepancy between two output heads [26]. Berthelot *et al.* proposed a MixMatch framework that integrated multiple components to improve the model performance using unlabeled data. They first utilized the mix-up approach [57] to transform original unlabeled samples into K augmented versions and employed a sharpening algorithm to minimize the entropy of predictions.

2.2.1 Leveraging Unlabeled Wearable Data. Semi-supervised learning methods have also been applied in mobile and wearable device studies [3, 31, 42], where models learned representations from massive unlabeled data and used a small amount of labeled data to train the supervised prediction models. Ballinger *et al.* applied a semi-supervised auto-encoder pre-training method [11] and used physiological data collected from wrist-wearable devices to detect symptoms of cardiovascular disease. Their semi-supervised model improved AUC scores for three out of four types of symptoms, and the highest improvement rate was 10.5% in detecting high cholesterol. In momentary stress detection, Maxhuni *et al.* used a self-training tree model in combining unlabeled wearable sensor-based physiological data with labeled data [31]. Compared to the supervised learning method, their binary stress detection performance in the F1 score boosted from 66.0% to 70.0%.

Besides the SOTA works in computer vision, leveraging unlabeled samples with contrastive constraints and consistency regularization have also been examined on wearable sensor data for human activities recognition (HAR) tasks [2, 21, 28, 36, 44, 45] and emotion classification [38, 49]. For example, Saeed *et al.* designed a pre-training task of distinguishing various transformations on the original unlabeled data, then transferring the

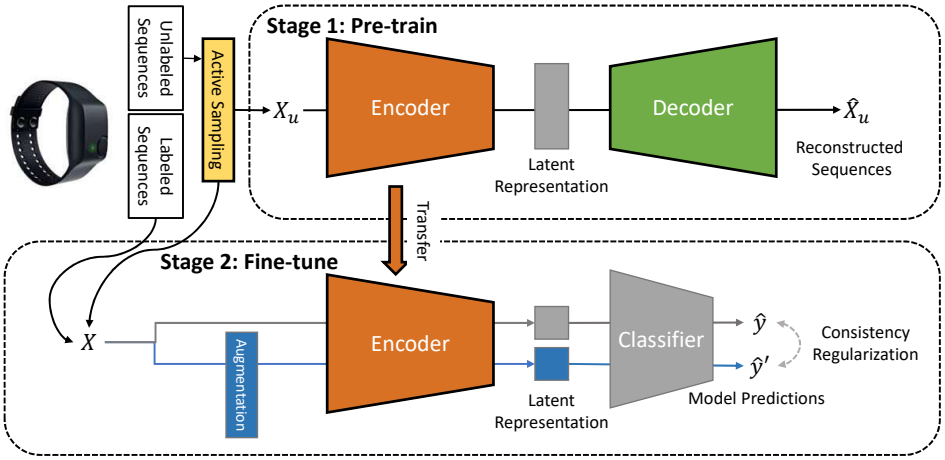

Fig. 1. The overall structure of the designed semi-supervised sequence learning framework for stress estimation. The framework consists of two stages of training with an active sampling module for selecting unlabeled samples to speed up the semi-supervised training process.

learned representations to supervised learning with labeled data. Similarly, Chi *et al.* pre-trained an encoder by classifying the data augmentations using unlabeled data and transferred the learned representations for HAR [45]. Liu *et al.* proposed a semi-supervised SimCLR[9]-based framework named SemiC-HAR to leverage the unlabeled data for HAR. [28] Jain *et al.* designed a contrastive constraint on the different temporal positions of unlabeled data collected from a multi-sensor accelerometer system. These methods achieved state-of-the-art performances in multiple applications with the prerequisite that the labeled and unlabeled data are in the same distribution space. However, for in-the-wild stress detection tasks, unlabeled samples can be in different distributions from the labeled samples even when those samples were collected from the same subjects. Besides, data augmentation methods were not guaranteed to help the model learn robust parameters. Thus, in this work, we propose to improve these methods by actively selecting unlabeled samples based on the latent distribution for semi-supervised learning. Also, we designed an averaged consistency regularization method in case of the improper noise introduced by augmentations.

3 METHODS

This section introduces our proposed semi-supervised learning method for leveraging all the data X, including both labeled X_l and unlabeled sequences X_u in detection stress labels y. A two-stage semi-supervised learning framework is proposed in this section. Moreover, we introduce a domain knowledge-based pseudo-annotating method to enrich the training data.

3.1 Semi-supervised Learning Framework

Figure 1 shows the overall framework of the proposed semi-supervised learning method. Using the actively sampled unlabeled data, the encoder, which is employed as the feature extractor, is pre-trained in an unsupervised manner with an auto-encoder structure. Further, the fine-tuning process involves both the labeled and unlabeled data. The consistency regularization with sequential data augmentation approaches is introduced to improve the robustness of the model. Following countless proven studies on modeling sequential physiological data [35], we use a one-dimension convolutional neural network (1D CNN) to extract information directly from the raw

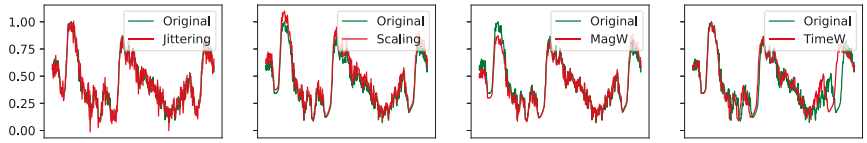

Fig. 2. Examples of data augmentation on a sequential sample

data. Thus, there is no feature extraction procedure required in this study. Additionally, unlike previous relevant HAR frameworks, such as SemiC-HAR, our framework does not incorporate pre-trained supervised model-based pseudo-label annotation, due to the challenging nature of the stress detection task. Instead, besides the basic two-stage proposed framework, we propose a novel unsupervised representation-based active sampling method for selecting unlabeled samples, which is introduced in Section 3.1.3.

3.1.1 Stage 1: Unsupervised Pre-training with Auto-encoder. Typically, a supervised learning inference can be equalized in a close form of:

$$\hat{y} = h(f(X)) \tag{1}$$

f is an encoder for extracting the features from signal X, and h is the classifier for predicting output $\hat{y}$. Without the stress labels, the optimization of h becomes difficult as lacking supervision for $\hat{y}$. However, learning robust parameters in the encoder f is feasible with the information contained in the unlabeled physiological data. In this study, we applied the 1D CNN-based structures to pre-train the model through the unlabeled samples.

The structure of an auto-encoder can seem as follows:

$$\hat{X}_u = g(f(X_u)) \tag{2}$$

Same as the supervised learning structure, f represents the encoder. With the extracted representations from X_u, the auto-encoder aims to reconstruct the input signal X_u as $\hat{X}_u$ by decoder g on top of the learned features $f(x_u)$. Thus, the objective function of the auto-encoder can be defined as the mean-square-error (MSE) loss between X_u and $\hat{X_u}$:

$$\mathcal{L}_{ae} = \|X_u - \hat{X_u}\|_2^2 \tag{3}$$

By optimizing the $\mathcal{L}_{ae}$, the encoder learns the weights of extracting informative features that represent the input signal. After training the auto-encoder, the parameters learned in the f are transferred to the model of stage 2 as the initialization of the encoder.

3.1.2 Stage 2: Semi-supervised Fine-tuning with Consistency Regularization. Consistency training methods regularize model predictions to be invariant to slight noise applied to input [32, 51]. The theoretical foundation of this method is that a robust machine learning model should be able to tolerate any slight noise in an input example. For example, when inputting a data sequence and its augmented sequence into a robust model, the outputs of those two input examples should be the same. Since there were plenty of unlabeled sequences in our datasets, ideally, the concept of consistency regularization could bring robustness to our model.

In our study, inspired by [51], we conducted consistency training combined with the augmented data for time-series data. As illustrated in [52], the effectiveness of data augmentation on sequential data is affected by multiple factors. Thus, we designed a multi-head consistency regularization to avoid the risk of learning ill-state representations with inappropriate augmentations. We randomly generated M augmented sequences using labeled/unlabeled samples in the same training batches. The consistency loss aims to regularize the similarity of predictions from the original data and the augmented samples. For example, since our task was to estimate stress status in binary classification, the supervised loss was designed as a cross-entropy loss. We applied the

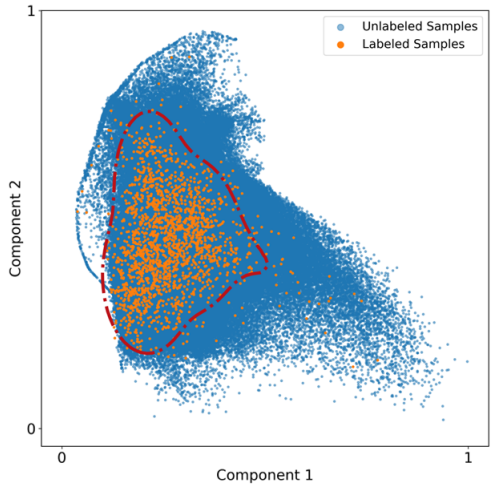

Fig. 3. Latent space PCA-based low dimension mapping visualization. The representations of labeled samples are highlighted in orange color. The red dashed line indicates the boundary of selecting unlabeled samples (blue). Example visualization in the SMILE dataset with four Gaussian mixture components.

Kullback-Leibler divergence loss as our designed consistency loss. To present the method in the formula, the final objective function of training stage 2 with the consistency regularization method is:

$$L = L_{CE}(X_l, y) + \frac{\alpha}{M}\sum_{m=1}^{M} L_{KL}(p(\hat{y}|X), p(\hat{\bar{y}}|\bar{X}^m)), \quad X = \{X_l \cup X_u\} \tag{4}$$

where $\bar{X}_l$ is the augmented labeled sequence, and $\bar{X}_u$ is the augmented unlabeled sequence. The probability $p(y|x)$ indicates the likelihood of getting model results with given data x. In our case of classification, $p(y|x)$ is the sigmoid output for binary classification. α controls the weights of the consistency regularization. The supervised consistency regularization coefficient α is set with a ramping-up function $w(t)$ to avoid noisy distortion in the early training stage.

$$\alpha = w(t) = c \cdot e^{(\min(\frac{epoch}{E_{warmup}},1)-1)^2} \tag{5}$$

In the above equation, *epoch* is the ongoing training epoch number, and E_{warmup} indicates the epoch number needed to warm up the consistency training. Here we set c to 1 and E_{warmup} to 50.

Data Augmentation. To perform the consistency regularization, we adopted four types of data augmentation techniques for time-series data from [47], including jittering, scaling, time warping, and magnitude warping. Jittering (J) adds tiny Gaussian noise to the original signals. For scaling (S), the original signals are scaled by generated Gaussian random numbers ($\mathcal{N} \sim (1, 0.05)$). Time warping (TW) perturbs the temporal characteristics of the data. The temporal locations of the samples are changed by smoothly distorting the time intervals between samples. Magnitude warping (MW) changes the magnitude of each sample by convoluting the data window with a smooth curve varying around one with a standard deviation of 0.05 ($\mathcal{N} \sim (1, 0.05)$). The essence of these methods is adding a small amount of noise to time-series data so that the trained model will be robust. Figure 2 shows an example of different DA methods on a sequence of electrodermal activity data. The green lines are the original signal, and the red lines represent the data generated using four different DA methods.

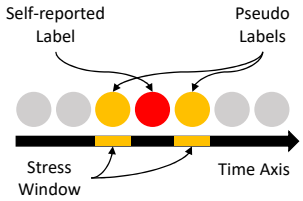

Fig. 4. The stress pseudo annotation approach. The equal value labels are assigned to the neighbor time points within symmetric windows of the self-reported time point.

3.1.3 Active Unlabeled Sample Selection. Wearable physiological signals collected in the wild contain noise. In addition, human physiological signals can vary widely in different states. These noises and uncertainty can cause significant distribution differences between labeled and unlabeled data samples. The distribution of the unlabeled data is usually more comprehensive, and a significant fraction of them might even distribute differently from the labeled data.

To reduce the influence of noise and unlabeled samples with different distributions to the model, we propose an active unlabeled sample selection method. We first trained an auto-encoder structure as in 3.1.1 with labeled data, then clustered all labeled samples in latent space low-dimension representation using a Gaussian mixture model (GMM). After analyzing the elbow points of both the Akaike and Bayesian information criteria, we fixed the number of Gaussian components as K. Then, we used the trained encoder to infer the latent representations of all the unlabeled samples as $f(x_u)$. The negative log-likelihood of each unlabeled sample, which is the probability of the observed data under the trained GMM model, can be calculated via the following equation:

$$\ell(\mathbf{x_u}) = -\log\left(\sum_{m=1}^{K} \gamma_m \phi(\mathbf{f}(\mathbf{x_u})|\mu_m, \Sigma_m)\right) \tag{6}$$

where γ represents the weight mixture component, μ and Σ are the learned mean value and co-variance of the corresponding Gaussian component.

Then, we selected the unlabeled samples that most possibly obey the similar distributions of labeled samples based on the calculated negative log-likelihood values (NLL). The smaller the NLL, the more similar the sample distributed as labeled data. Figure 3 shows the reduced-dimensional visualization of latent space representations, and the red dash line indicates the example of sampling boundary by negative log-likelihood levels across the whole dataset. Under this scenario, we only focus on the unlabeled data within the red boundary in the semi-supervised learning framework.

3.2 Pseudo Annotation

Based on the fact that stress in human daily life may not change rapidly, stress status usually takes minutes to be relieved [20]. While in real-world studies, self-reported stress annotation only represents the subjective stress status of study participants at a certain time point. To enrich the number of labels with the slow-changing attribute of human stress, we applied a time window-based pseudo annotating strategy as illustrated in Figure 4. A time window t was applied before and after the time point where study subjects reported their stress status y, and the time points within t are annotated with the same label value of y. The window length t was treated as a hyperparameter that requires fine-tuning while training.

Table 1. Meta information about datasets used for evaluation

Dataset	SMILE	TILES
# of Labeled Sequences	2494	1229
# of Unlabeled Sequences	480,000	370,000
# of Classes	2	2
Used Modality	ECG, GSR, ACC, ST	ECG
# of Participants	45	212

4 EXPERIMENTAL EVALUATION

This section introduces experimental settings such as datasets, model structures, training hyperparameters, etc. The baseline models and the reproduced SOTA methods are also described. Then, the evaluation results of the proposed method are listed for various methods and datasets.

4.1 Datasets

We describe two datasets we used to evaluate our methods. The meta-information of the datasets can be found in Table 1.

4.1.1 Dataset I: SMILE. Wearable sensor and self-report data were collected from 45 healthy participants (39 females and 6 males) for 390 days. The average age of participants was 24.5 years old, with a standard deviation of 3.0 years. Participants contributed to an average of 8.7 days of data, with a minimum of 5 days and a maximum of 9 days. Two types of wearable sensors were used for data collection [40]. One was a wrist-worn device (Chillband, IMEC, Belgium) designed for the measurement of skin conductance (SC), ST, and acceleration data (ACC). The SC was sampled at 256 Hz, ST at 1 Hz, and ACC at 32 Hz. Participants wore the sensor for the entire testing period but could take it off during the night and while taking a shower or during vigorous activities. The second sensor was a chest patch (Health Patch, IMEC, Belgium) to measure ECG and ACC. It contains a sensor node designed to monitor ECG at 256 Hz and ACC at 32 Hz continuously throughout the study period. Participants could remove the patch while showering or before doing intense exercises.

In addition to the physiological data collected by sensors, participants received notifications on their mobile phones to report their momentary stress levels 10 times per day, spaced out roughly 90 minutes apart for eight consecutive days. In total, 2494 stress labels were collected across all participants (80% compliance). The stress scale ranged from 1 ("not at all") to 7 ("Extreme"). In 45% of the cases, participants reported that they were not under stress, while in only 2% of the cases did they report that they were under extreme stress.

Data Processing: In this work, we focused on the physiological signals of ECG and GSR in modeling stress status. We preprocessed the ECG data with a high-pass Butterworth filter with a cutoff frequency of 0.5 Hz and an order number of 5. Also, a powerline filtering for the white noise at 50Hz was applied after the Butterworth filter. For the raw GSR data in the SMILE data, we decomposed the signal into phasic and tonic components, and only the tonic components were used as the model input in order to preserve the long-term temporal information and avoid introducing noises to the model. The implementation of the aforementioned procedures was based on the NeuroKit2 Python library [29]. The ECG and GSR signals were segmented into one-minute windows and aligned based on the timestamps. Around 480,000 bimodal sequential inputs were constructed with 2494 annotations. Further, for the stress labels, we binarized the stress levels by categorizing stress level 1 as a class of "non-stressed" (45%) and levels 2-7 as the "stressed" class (55%).

4.1.2 Dataset II: TILES. Tracking Individual Performance with Sensors (TILES) is a multi-modal data set for the analysis of stress, task performance, behavior, and other factors to professionals engaged in a high-stress

workplace environment [33]. The dataset was collected from 212 participants for 10 weeks. In this work, we leveraged the ECG signals collected by the chest-worn OMSignal smart garments, which were not collected in a strictly continuous manner. At 5-minute intervals, the sensor collected ECG signals for fifteen seconds at a sampling rate of 250 Hz for the participants. Regarding the stress labels, participants annotated stress levels through multiple 5-point scale questions.

Note that prior research has been conducted on this dataset. For example, Gaballah *et al.* leveraged TILES audio and physiological data with a bidirectional LSTM network and inferred stress labels in a binary classification task with an F1-score of 0.64[14].

Data Processing: Since our interest lies in leveraging physiological data collected from wearable sensors, only the ECG data were processed and utilized. Similar to the preprocessing procedures for ECG for the SMILE dataset, we applied a high-pass Butterworth filter with a cutoff frequency of 0.5 Hz and an order number of 5; also, a 50 Hz powerline filter was applied for filtering out the white noise. However, since the ECG was not collected continuously in the TILES dataset, we used 15-second data sequences every 5 minutes, and the alignment of ECG and stress annotations are also in a resolution of 5 minutes. Following the stress label processing procedures in [14], We calculated the z-scores of stress levels for each individual, considering the subjective variability, and then divided them into two classes, class 0 (non-stressed, z-score below the average) and class 1 (stressed, z-score above the average). Overall, around 370,000 sequential samples were processed with 600 stressed labels and 629 non-stressed labels.

4.1.3 Training & Testing Sets Split. To fairly leverage the data from subjects, we applied a subject-independent cross-validation setting in training and testing sets split. Five folds are split evenly for both SMILE and TILES datasets. For example, for the SMILE dataset, on each split, data from 36 subjects are used as the training set, and the rest data is employed as the testing set for performance evaluation.

4.2 Model Structures, and Hyperparameters

To ensure the reproductivity of the proposed method, we describe the model structures and hyperparameters of the proposed method and the reproduced semi-supervised learning SOTA methods.

4.2.1 Proposed Method. To implement the proposed structure, we utilized the model structures including, *ECG encoder*, *ECG decoder*, *GSR encoder*, *GSR encoder*, and *classifier*.

ECG Encoder & Decoder. We used the same 1D CNN structure in extracting information from ECG sequences for both SMILE and TILES data as the sampling frequencies of measuring ECG are close. The ECG encoder structure consists of 5 layers of CNN layers with kernel sizes of [8, 6, 5, 3, 3] and samples the channels of signal from 1 into [16, 32, 64, 128, 256] respectively. We used an average pooling layer at the output of the encoder and obtain vectors with a length of 256 for the classifier.

The structure of the ECG decoder was symmetric as that of the ECG encoder. Totally 5 layers of the 1D transposed convolutional layers (1D TranCNN) function to reconstruct the input signal directly from the output of the encoder. Thus, the kernel size and the out-channel number were [3, 3, 5, 6, 8] and [128, 64, 32, 16, 1], respectively.

GSR Encoder & Decoder. The GSR encoder and decoder were used to extract latent representations for the SMILE dataset. The encoder structure contained 3 layers of CNN layers with kernel sizes and out-channel dimensions of [8, 5, 3] and [16, 32, 64], respectively. Symmetrically, the kernel sizes and the out-channel dimensions for the TranCNN layers of the decoder structures were [3, 5, 8] and [32, 16, 1], respectively.

Classifier. As the design of the encoder structures, we obtained features in dimensions 256 and 64 for ECG and GSR, respectively. Thus, the input dimension of the classifier for the SMILE dataset was 320; and the dimension of

the TILES dataset was 256. One embedding layer of 512 dimensions connected the input layers for both datasets, and an output layer for class number 2 is employed.

4.2.2 Reproduced SOTA Semi-supervised Learning Methods. To conduct an empirical comparison between our proposed method and the SOTA methods on the wearable-based stress detection task, we reproduced the following methods that have been proven in computer vision tasks: two consistency regularization-based SOTA methods such as Π-model [26] and virtual adversarial training (VAT) [32] and two hybrid semi-supervised learning methods including interpolation consistency training (ICT) [50], MixMatch [5], and FixMatch [41]. Also, we conducted experiments on the proven methods of leveraging unlabeled samples on wearable data for HAR, e.g., SemiC-HAR [28] and SelfHAR [45], as reproduced SOTAs for comparison. See the details about the SOTA semi-supervised learning methods below.

Π-model [26]: The Π-model operated two different transformation for an unlabeled input x_u, to form x'_u and x''_u so that the model predicted y'_u and y''_u. Then the model constrains the consistency of the two results. We implemented Π-model with two different DA approaches in section 3.1.2 randomly to form different input data transformations for each training sample. The mean squared error was used as the consistency loss. Based on the model performance in model training, we adjusted ramping up epoch in equation (5) as 60.

VAT [32]: This algorithm constrained the consistency of a signal and its transformation with additive noise, the trainable adversarial perturbation r. The perturbation r was trainable, which was constrained by coefficient $\xi = 1 \times 10^{-6}$ to avoid gradients explosion in our implementation. We allowed 5 iterations for each sample in a single epoch to update the parameter of r. Based on the model performance in model training, we adjusted to ramping up epoch in equation (5) as 30 with a coefficient c of 0.3.

ICT [50]: The ICT algorithm used the mix-up method, which summed the original unlabeled data to generate the augmented samples. In our implementation, the mix-up coefficient was set as 0.2, which means we summed up $0.8 \cdot x_u^1$ and $0.2 \cdot x_u^2$ as a new signal as x'_u. Then, the model optimized the discrepancy between the prediction y'_u and $\{0.8 \cdot y_u^1 + 0.2 \cdot y_u^2\}$. Also, we reproduced the average teacher strategy [46] with an updating factor of 0.999. We adjusted the ramping up epoch in equation (5) as 20 with a coefficient c of 80.

MixMatch [5]: The MixMatch approach combined multiple prior techniques, such as consistency regularization, entropy minimization, and mix-up DA approach, to serve as a semi-supervised learning framework. Similarly to ICT, we also reproduced the mix-up approach in the MixMatch algorithm with a mix-up coefficient of 0.2. In the steps of sharpening prediction and reducing model entropy, we set the averaging bag size as 3 for each sample with a normalizing temperature of 0.5. The ramping-up epoch in equation (5) was set to 30 with a coefficient c of 100.

FixMatch [41]: The FixMatch approach relied on the consistency of the outputs from the weakly-augmented samples and strongly-augmented samples. We set the jittering and scaling methods as the weak augmentations; while the TimeW and MagW as the strong augmentations. Based on the model confidence from the weakly-augmented predictions, pseudo labels were annotated to enrich the training set. We set the threshold of pseudo labeling as a threshold of 0.95 to the model output after softmax. The ramping-up epoch in equation (5) was set to 20 with a coefficient c of 80.

SelfHAR [45]: The SelfHAR is a teacher-student-based representation learning that pre-trains the model encoder with a pre-task of recognizing various data transformations applied to the original signal. We reproduced this approach based on the data augmentation methods we introduced in section 3.1.2, and the unlabeled samples were selected from the teacher model with a confidence threshold of 0.9. With 100 epochs of pre-training, the accuracy rate of distinguishing augmentation methods achieved over 98%. Then, the pre-trained weights in the encoder served as the initial encoder weights in supervised learning. We fine-tuned the model for the stress detection task in 50 epochs with a learning rate of 1e-3.

SemiC-HAR [28]: The SemiC-HAR presents a 4-stage framework that involves supervised training, self-labeling, contrastive learning with selected samples, and fine-tuning. The framework used the labeled data to train a supervised learning model and annotate the unlabeled samples with the model logits as the likelihood scores from the trained supervised model. Then, the framework pre-trained the encoder using SimCLR [9] with samples selected by the prediction confidence, which was set to 0.9 following the original study. Finally, the pre-trained encoder was fine-tuned with our stress prediction task in 50 epochs with a learning rate of 1e-3.

4.3 Performance Evaluation

In the performance evaluation section, we aim to answer the following research questions on the datasets introduced in section 4.1:

- ***Q1. What is the baseline performance of stress prediction in the wild?***
- ***Q2. Do all the semi-supervised learning components contribute to the performance?***
- ***Q3. Is active sampling unlabeled data helpful in stress detection?***
- ***Q4. How do well-proven semi-supervised learning methods from other domains work in stress detection?***

To understand ***Q1***, we designed two baseline approaches including a random guessing baseline, a major class baseline, and a purely supervised 1D CNN baseline. Random baseline is the method that assigns labels to test instances according to the class probabilities in the training set [6]. For example, in the SMILE dataset, the probability of class 0 $p(y = 0) = 0.45$, we assigned the instances in the test set as class 0 with the probability of 0.45. Major class baseline means we predict all the samples from the evaluation dataset to be in the major class, e.g., predicting all "stressed" for the SMILE dataset and all "non-stressed" for the TILES dataset. Besides, we examined the model performances of using pseudo labels proposed in section 3.2 along with the baseline methods, as this method is straightforward and intuitive. The tuning of the time-window length is discussed with the ablation studies as section 6.1.

We answer ***Q2*** by experiments utilizing the semi-supervised methods, including auto-encoder pre-training and consistency regularization, in individual and combined manners. Further, by testing the performances of whether conducting the active sampling unlabeled samples, we explore the question ***Q3***. Moreover, we examined and compared the performances of the reproduced methods in section 4.2.2 to answer ***Q4***. Note that, the experiments in ***Q2, Q3***, and ***Q4*** are based on the training set that has been expanded by pseudo labels.

5 RESULTS

Table 2 shows the results of evaluations in the average accuracy rate and macro scores with standard deviations. According to model performances, we summarize our results as follows:

Q1 (Baseline stress prediction performance): According to the performance comparison between the random baseline and the 1D CNN baseline, we found that stress detection in the wild is intuitively challenging. Even though statistical differences (paired t-test, $p < 0.05$) were observed on both datasets for both accuracy and macro F1 scores, the supervised learning method output only outperformed the random guessing by a small margin. Also, we found that applying the time windows-based pseudo labels in supervised learning improved the model performance significantly (paired t-test, $p < 0.05$).

Q2 (Contributions of Semi-supervised learning): In Table 2, we listed the performance of using the semi-supervised framework on top of the pseudo labels. For both auto-encoder pre-training and consistency learning approaches, applying them individually or together boosted the model performance compared to using pseudo labels only by clear margins. Also, the results suggested the consistency regularization method achieved higher performances than the auto-encoder pre-training method; while the combination of both methods provided the best performances on both datasets (ANOVA, Tukey, $p < 0.05$.).

Table 2. Model Performances of 5-fold cross-validation using different methods (macro F1 score). AE: Auto-encoder pre-training, CR: consistency regularization, AS: active sampling. The semi-supervised learning methods (all below pseudo labels) used pseudo labels. **Bold** represents the statistical differences (ANOVA, Tukey, $p < 0.05$)

	SMILE		TILES	
	ACC	F1-macro	ACC	F1-macro
Baseline: Random	49.61 (1.21)	46.33 (0.84)	50.67 (1.15)	46.71 (0.96)
Baseline: Major Class	55.00 (0.47)	35.50 (0.62)	50.45 (0.82)	32.93 (0.77)
Baseline: 1D CNN	53.34 (3.39)	52.98 (4.61)	52.79 (2.71)	50.79 (3.21)
Pseudo Labels	57.88 (2.26)	57.59 (2.33)	54.63 (2.74)	55.06 (2.97)
Π-model [26]	60.32 (2.25)	60.51 (2.12)	57.73 (2.49)	57.78 (2.25)
VAT [32]	62.43 (2.53)	61.98 (2.56)	58.26 (2.33)	58.37 (2.49)
ICT [50]	59.31 (3.09)	58.92 (3.14)	55.59 (3.28)	55.77 (3.21)
MixMatch [5]	59.89 (2.59)	59.81 (2.44)	56.41 (2.57)	56.09 (2.87)
FixMatch [41]	**62.72 (1.87)**	62.35 (2.27)	57.98 (2.14)	58.07 (2.26)
SelfHAR [45]	58.26 (2.17)	57.89 (2.97)	55.73 (2.55)	55.24 (2.71)
SemiC-HAR [28]	56.78 (3.66)	54.34 (4.01)	54.67 (2.42)	54.88 (2.57)
AE	60.35 (2.37)	60.16 (2.22)	57.14 (1.66)	57.29 (1.80)
CR	62.32 (2.08)	62.21 (2.15)	58.47 (1.84)	58.55 (2.17)
AE + CR	**62.83 (1.94)**	**62.79 (1.95)**	58.99 (1.73)	59.04 (1.84)
AS + AE	61.08 (2.19)	60.89 (2.31)	57.77 (1.42)	58.02 (2.30)
AS + CR	**63.17 (1.85)**	**62.91 (2.00)**	59.01 (1.78)	59.16 (2.01)
AS + AE + CR	**63.44 (2.05)**	**63.21 (1.77)**	**59.64 (1.61)**	**59.57 (1.54)**

Q3 (Contributions of Active Sampling): The last 3 rows in Table 2 show the performance of semi-supervised learning approaches with active sampled unlabeled data. Compared to the results without applying the active sampling techniques, active sampling significantly improved the model performance for both auto-encoder pre-training and consistency regularization on both the SMILE and TILES datasets (paired t-test, $p < 0.05$). Nevertheless, we did not observe statistical differences when comparing the results of using both semi-supervised methods on the SMILE dataset even with higher averaged performances (paired t-test, $p > 0.05$).
Q4 (Comparison with SOTA semi-supervised learning methods): The performances from 5 reproductions of well-proven semi-supervised methods all achieved improvements on top of the pseudo-labeling method. On the SMILE dataset, FixMatch achieved the best performance in both accuracy and F1 score; whereas VAT performed the best on TILES. However, ICT and MixMatch, which were developed on top of the mix-up augmentation approach showed lower margins of improvement compared to other methods. When comparing these reproduced methods with our proposed framework, our method showed statistically high macro scores than those with and without an active sampling approach applied (ANOVA, Tukey, $p < 0.05$).

6 ABLATION STUDIES

This section covers the ablation studies of essential hyperparameter tuning in this study.

6.1 Window Lengths for Pseudo Annotations

As shown in the results section, enriching the training set can help improve the test performance in stress detection. However, the proper length of the window of applying pseudo labels is the essential hyperparameter

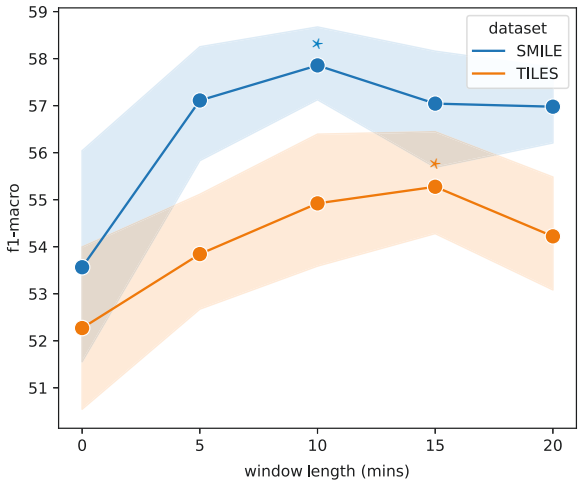

Fig. 5. The performance plot of using various lengths of time windows for generating pseudo labels. The * marker indicates the window length we selected for the other experiments in this study.

to determine. Larger window sizes inject more samples into the training set while introducing the risk of biased samples at the same time. This section shows experiments with time window tuning on both the SMILE and TILES datasets. Figure 5 shows the performances of applying different time windows in generating pseudo labels. For both the SMILE and TILES datasets, we observed the trends that the model performance first increased and then decreased with longer window sizes. Based on these results, we fixed the window sizes to be 10 and 15 minutes for the SMILE and TILES datasets, respectively.

6.2 Clustering Methods and Volumes of Actively Sampled Unlabeled Data

We propose an active sampling method for selecting unlabeled in this study. To conduct the most effective and efficient training process, we may select the least amount of data to get the highest performance as the ideal cases. Further, since our proposed method relied on the cluster of the latent representations, we also explore another clustering algorithm, k-Means (k=6), to compare with the GMM method proposed in section 3.1.3. Moreover, to verify the effectiveness of active unlabeled sample selection, we also evaluated the pre-trained models with randomly sampled data. We tune the volume of the selected unlabeled samples by conducting experiments on both datasets in the auto-encoder pre-training task. Figure 6 shows the performances of using different portions of the unlabeled samples while tuning. For both SMILE and TILES data, the active sampling method outperformed the random sampling baseline. Also, with active sampled unlabeled data, the semi-supervised learning algorithm achieved the highest performances without leveraging all the unlabeled samples. In addition, based on our ablation studies, we selected GMM as the clustering algorithm instead of the kMeans as GMM provided higher macro F1-scores with fewer unlabeled samples. Based on the results, we selected 40% of the unlabeled samples for both the SMILE and TILES datasets using GMM.

6.3 Physiological Modality Selection

As introduced in section 4.1, the datasets of both SMILE and TILES contain multiple physiological modalities. Besides the modalities we leveraged in the experiments, ST and ACC are available in SMILE, and minute-to-minute features such as heart rate, step count, and sleep status are available with the Fitbit wristband in the TILES dataset. Considering the complexity of the model, we desire to achieve the best results with the least amount of modalities. Thus, we conducted ablation experiments on the selection of modalities. For example, Table 3

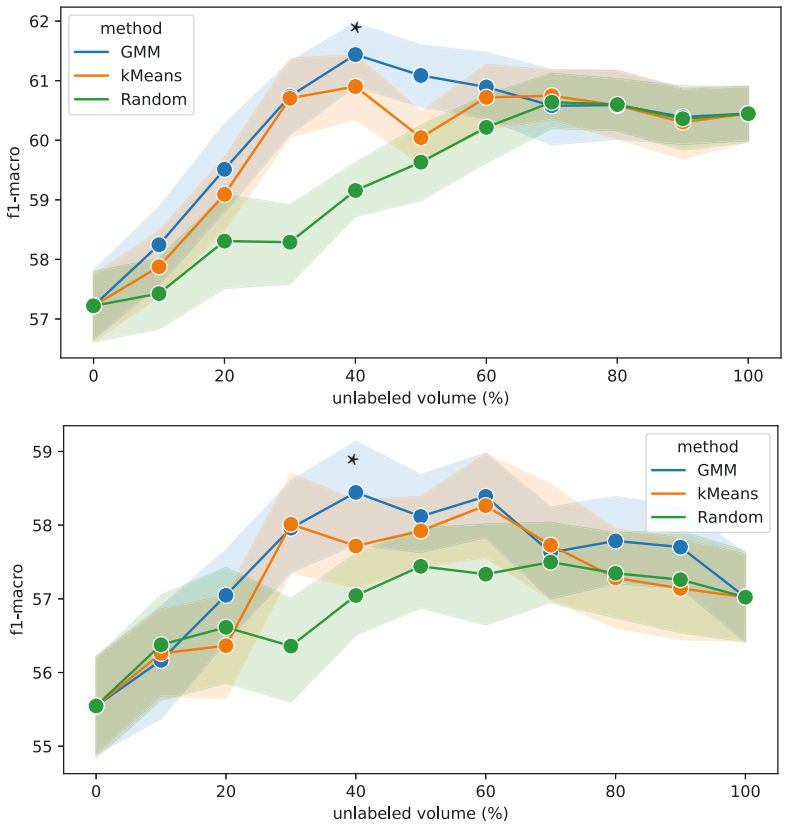

Fig. 6. Model performances versus the volumes of the selected unlabeled samples for SMILE (upper) and TILES (lower) datasets. Auto-encoder pre-training method was used in tuning the selected volumes. And methods including GMM, kMeans, and random sampling were compared. The * marker indicates the volume of the unlabeled samples we selected for the other experiments in this study.

shows the average model performance of supervised learning 1D-CNN on the SMILE dataset using different combinations of modalities. From the table, we found the supervised model did not learn useful information in detecting stress using modalities such as ST and ACC. The combination of ECG and GSR modalities provided the best results. Our results are consistent with previous findings in the literature. For instance, GSR is a tonic and phasic electric signature on the skin that changes when a person experiences stress or emotional arousal as a result of sweat gland activation [7]. ECG reflects human heart activity, including heart rate variability, which decreases under stress due to sympathetic nervous system activation [48]. Researchers have also demonstrated that utilizing and combining multiple sensor data in multimodal methods, enabled by these physiological signals, can provide additional benefits [15, 17, 55].

On the TILES dataset, we examined the supervised 1D-CNN performance with combinations of (1) ECG, (2) Fitbit features, and (3) ECG + Fitbit features. We observed the supervised performances for these three combinations to be 50.79 (3.21), 48.52 (3.45), and 50.90 (3.27), respectively. Although the combination of ECG and Fitbit showed slightly higher average performance than using ECG only. We did not find any significant

Table 3. Supervised 1D-CNN performances using different modalities on the SMILE dataset. Metric: average macro F1 score

ECG	GSR	ST	ACC	F1-macro
✓				50.73 (4.06)
	✓			52.33 (3.89)
		✓		35.50 (0.55)
			✓	43.64 (4.11)
✓	✓			**52.98 (4.61)**
✓		✓		47.21 (5.04)
✓			✓	49.54 (4.70)
	✓	✓		50.86 (3.47)
	✓		✓	51.57 (3.82)
		✓	✓	40.05 (4.39)
✓	✓	✓		51.58 (4.83)
✓	✓		✓	52.76 (3.61)
	✓	✓	✓	51.97 (4.26)
✓	✓	✓	✓	52.19 (4.02)

Table 4. The performance in weighted F1 score of applying the proposed semi-supervised method on HAR tasks.

	HHAR	Motion Sense	WISDM	UCI-HAR
Supervised	0.7924	0.9173	0.8978	0.8990
SelfHAR	0.7739	0.9312	0.8809	0.8927
SemiC-HAR	0.8510	0.9393	0.9006	0.9264
Ours	0.8379	0.9355	0.9231	0.9270

difference (paired t-test, $p > 0.05$). Considering the results of the ablation experiments, we selected physiological modalities of ECG and GSR for the SMILE dataset, and ECG only for the TILES dataset.

6.4 Evaluation in HAR Datasets

We proposed and evaluated a semi-supervised learning framework in stress detection; however, auto-encoder pre-training and consistency regularization are task-agnostic. Thus, as an ablation study, we tested the capability of the proposed method in the HAR tasks. Following the settings in [28, 45], we applied our method to four HAR datasets including the Motion Sense [30], HHAR [43], UCI HAR, and WISDM[25] datasets, which contain motion-related sensing data collected from portable devices and labels of different activities. The raw datasets are pre-processed using the released source codes in [45], which generates pairs of sequential samples and activity labels. To construct the sparsely annotated datasets, we followed the same strategy as in [28], which was to mask out 90% of the samples as unlabeled samples and conduct the experiments on the rest 10% of the samples. We randomly selected 10% of the labels in this ablation study. The weighted F1 score was used as the evaluation metric. Table 4 shows the evaluation results of the conducted HAR ablation study. We observed that compared to the supervised model, our method improved the performance using the unlabeled samples. Also, our method achieved competitive performance compared to SOTAs. Note that in the HAR experiments, the proposed active sampling method did not contribute to the unlabeled sample selection, as the distribution shifts between the labeled and unlabeled samples were not significant. In addition, we did not annotate the pseudo labels as in section 3.2 as activities can change more rapidly than stress levels.

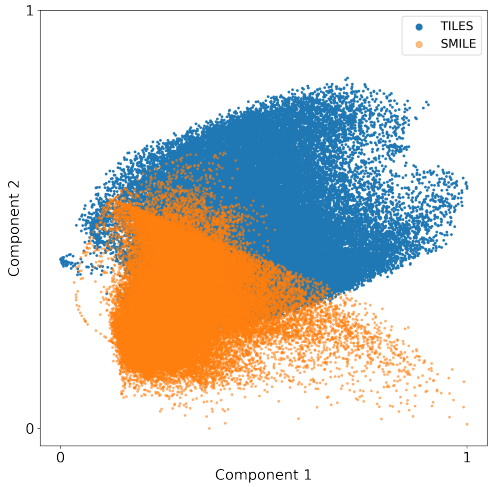

Fig. 7. The visualization of the latent space representation for ECG signals in the SMILE and TILES datasets.

6.5 Cross-dataset Evaluation

To further evaluate the cross-dataset robustness of the proposed method, we also explored the possibility of merging datasets with similar signals to enlarge the volumes of available annotations. For example, both the SMILE and TILES datasets contain in-wild stress labels and ECG signals. We conducted an ablation study of cross-dataset evaluation where we trained a model on the SMILE dataset but tested it on the TILES dataset and vice-versa, to examine the cross-dataset robustness. To adjust ECG signals in both datasets, we resampled the ECG signal from the SMILE dataset to 250 Hz, and truncated the samples in the SMILE dataset to a length of 15 seconds. To simplify the evaluation, we focused on two settings: (1) the supervised learning approach with original and pseudo labels, and (2) the semi-supervised learning approach with active sampling, auto-encoder pre-training, and consistency regularization. Table 5 shows the cross-dataset evaluation results in average macro F1 scores. From the table, we can observe substantial performance drops across different datasets. Multiple factors might result in these performance drops. For example, ECG data were collected using different devices (chest patch vs smart garment) and the location of measurement on the chest might be slightly different. Different scales were used to collect self-reported stress labels. Figure 7 shows the visualization of the latent space from SMILE and TILES datasets, where we can observe that the representations do not obey the same distribution. The results of this ablation study showed that merging different stress datasets can be challenging and requires additional approaches to tackle distribution drift issues.

7 DISCUSSION

In this section, we discuss semi-supervised learning algorithms in stress detection. Our results showed that the mix-up augmentation method, which has been well-proven in the other domain, did not perform well on our evaluating databases. Therefore, we discuss the insight into applying a mix-up approach for wearable data. Besides, we compare this work with the other works that utilized the same databases in the literature. Moreover, this section covers the biased label challenge we found during the experiments. Lastly, we conclude our discussion by summarizing the implication and limitations of this study.

Table 5. Cross-dataset evaluation performances in macro F1 scores on the SMILE and TILES datasets. The supervised method is based on the original labels and pseudo labels, and the semi-supervised methods cover active sampling, auto-encoder pre-training, and consistency regularization.

Training Set	Methods	Evalution Set	
		SMILE	TILES
SMILE	Supervised	54.36	48.73
	Semi-supervised	57.97	49.62
TILES	Supervised	47.33	55.06
	Semi-supervised	48.01	59.57

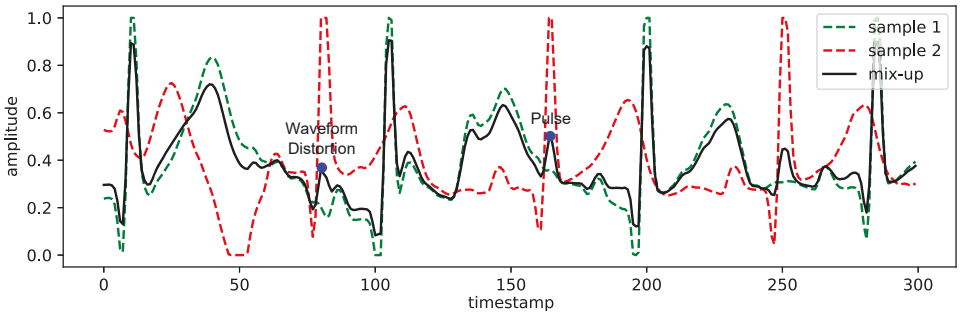

Fig. 8. An example of the mix-up augmentation method with a mix-up coefficient of 0.2.

7.1 Mix-up Augmentation in Sequential Physiological Data

The mix-up augmentation approach is well-applied in the computer vision domain and achieves promising results in semi-supervised learning studies. However, according to the results of the reproduced SOTA in section 4.3, we found the mix-up method did not benefit the semi-supervised learning in our datasets as observed in the computer vision domain. We believe the misalignment of the temporal information invalidated the mix-up augmentation. Figure 8 shows an example of mixing two ECG sequences. We observed that the mix-up method can generate excessive distortion of the ECG signal. For example, without alignment in the temporal positions, the difference in the signal amplitude between the peak and plateau can easily create pulses in the plateau region or distort the waveform at critical positions (e.g., QRS areas in ECG).

7.2 Comparison of Our Results and Prior Work that Used the Same Datasets

Several studies have been conducted for human momentary stress detection using the same databases. Even though there are differences in experimental settings, we compare our experiments and results with the studies in the literature. Since the raw data in SMILE are not publicly accessible at this moment, we focus on the comparison with the literature on TILES in this section.

Table 6 shows the comparison among the studies for stress detection tasks with the TILES dataset. Gaballah *et al.* extracted crafted features from audio, locations, and wearable sensors in 30-minute windows for consecutive 48 hours before the stress annotated time point [14]. They applied a bidirectional LSTM network to model the extracted feature sequences into stress status. Pimentel *et al.* [34] and Yang *et al.* [53] engineered features from wearable data, including ECG signal and features collected by Fitbit. Pimentel *et al.* [34] extracted high-level heart rate variability (HRV) features across a time window of 24 hours and predicted stress status with an SVM

Table 6. The table of comparing studies in literature for the TILES dataset

	Modalities	Sequence	feature	Acc	F1	F1-macro
Gaballah et al. [14]	Audio, Location, Wearable	48 hours	hand-crafted	~65	~65	-
Pimentel et al. [34]	Wearable (HRV)	24 hours	hand-crafted	62.4	65.7	
Yang et al. [53]	Wearable (HRV + Fitbit)	120 mins	hand-crafted	58.6	70.3	-
Ours	Wearable (ECG only)	15 seconds	raw data	59.64	61.49	59.57

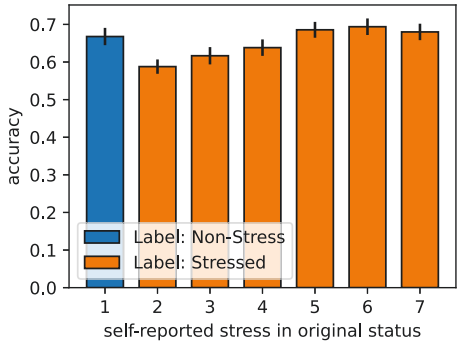

Fig. 9. The model performances in accuracy on the SMILE dataset versus the originally reported stress status before binarizing

model. Yang *et al.* [53] leveraged HRV and Fitbit features of every 5 minutes with a total of 120 minutes window (5 minutes $\times$ 24 time steps), and they developed the LSTM method to estimate the stress status for subjects. According to the evaluation results shown in the table, we found that leveraging the information in longer data sequences may contribute to the model performance in accuracy. Also, as indicated in [14, 52], multi-modal learning helped improve the performances in stress detection on the TILES dataset. Our method showed a similar level of performance in accuracy rate compared to [34, 53], with the information from remarkably shorted data sequences (15 seconds) of ECG data only.

7.3 Biases in Stress Labels

The stress labels are annotated by subjective self-reported questionnaires reported by study participants during the study period. Thus, it is highly possible that biases are introduced into the stress labels. Moreover, due to the participant-wise heterogeneity in stress perception, the stress labels can be biased across subjects when considering a generalized model. The biases in labels are reflected in the test accuracy. Figure 9 shows the model performance in accuracy for each individual stress status on the SMILE dataset. As mentioned in section 4.1, study participants reported their stress status in 7 different levels (1 - 7), and we binarized the stress labels by 1 vs. 2-7. From and figure, as well as according to the statistical test (ANOVA. Tukey, $p < 0.05$), we can observe lower performances in accuracy when the original stress status was reported as 2 and 3 compared to the original status ≥ 5. This phenomenon suggests that the model has difficulty distinguishing the features embodied by the physiological signal when the stress level is at the boundary point of whether or not the participant was stressed. In other words, the stress labels around the boundary point can be "biased" to the model as these samples were more challenged to be correctly classified by models. To avoid biases from these uncertain annotations, we propose discarding the labels in the boundary or using a different label strategy while binarizing the stress labels. For example, in our case, we could try non-stress vs. stressed in the original status of 1 vs. 4-7 on the SMILE dataset.

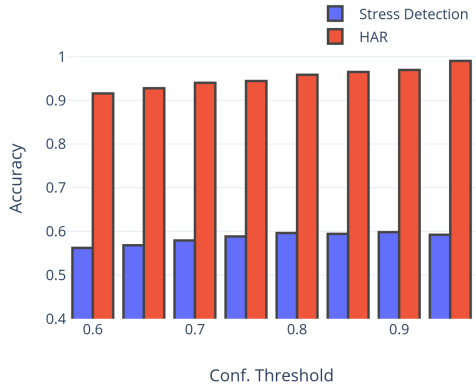

Fig. 10. A comparison of supervised learning-based pseudo-annotation generation thresholds for stress detection (SMILE) and Human Activity Recognition (MotionSense) tasks. Using the same selection thresholds, the accuracy of pseudo labels in the HAR study is substantially higher than that of the stress detection task.

7.4 Stress Detection versus Human Activity Recognition

We conducted a thorough evaluation of our proposed framework's performance on stress detection tasks. Additionally, we examined the model's performance on the HAR tasks, comparing our method with strong baselines such as SelfHAR and SemiC-HAR in our ablation studies. Interestingly, we found that the HAR-originated SemiC-HAR method performed competitively in HAR tasks, but exhibited weaker performance in stress detection tasks. This phenomenon might be related to the lack of robustness of the pseudo-annotation stage of SemiC-HAR, which relies on a pre-trained supervised classifier using model likelihood confidence scores as mentioned in Section 4.2.2.

In contrast to the HAR task, in-wild stress detection is more challenging, resulting in high incorrect annotation rates during the supervised learning-based pseudo-annotation step. For example, Figure 10 illustrates the relationship between the accuracy of supervised based pseudo labels and model likelihood confidence scores on the SMILE and Motion Sense datasets, respectively. The supervised accuracy on the Motion Sense dataset is substantially higher than the performance on the SMILE dataset. Furthermore, the supervised learning performance on the SMILE dataset suggests that even with a high selection threshold, the pseudo-labels based on supervised learning are noisier than those in the Motion Sense dataset. For instance, when using a threshold of 0.90, the accuracy in pseudo annotation on the Motion Sense dataset can be higher than 95%; whereas the accuracy on the SMILE dataset is lower than 60%. In addition, as revealed in Section 6.4, we did not observe a significant difference in the latent space representations between the labeled and unlabeled datasets in HAR tasks. Consequently, our unsupervised active sampling method did not fully function when performing the HAR tasks. Thus, although there are similarities, such as unsupervised pre-training and consistency regularization, between the methods, our proposed method outperforms the HAR-originated methods including SelfHAR and SemiC-HAR in stress detection.

7.5 Semi-supervised Learning and Self-supervised Learning

Semi-supervised learning and self-supervised learning have similarities in that they both aim to learn reasonable representations from unlabeled data. The main difference between them is whether they leverage labeled information in the process of representation learning or not. In this work, we also validated self-supervised learning methods, such as SelfHAR [45], which showed lower performances compared to the semi-supervised methods. This might be due to the risk of trivial representations learned by self-supervised learning. For example, the constraint of learning representation in SelfHAR is the disagreement among different augmented views;

however, the pre-task is not challenging enough as the model can achieve very high accuracy in distinguishing augmentations and might have learned trivial representations. Thus, we speculate that semi-supervised learning constrains the learning process more effectively with labeled samples in the training set.

7.6 Implication

In this study, we conducted empirical experiments on SOTA semi-supervised learning approaches in stress detection and human activity recognition tasks. Further, by proposing a semi-supervised learning framework with a novel hinge of active sampling method, we achieved even higher performance than the reproduced methods in the literature. We believe that our work can inspire future studies in wild stress or other human construct detection. Since label annotation is usually expensive, the lack of training samples becomes a common problem in this field. To the best of our knowledge, there are not many semi-supervised learning studies conducted on stress detection. Our work, on the other hand, significantly improves the performance of the model on top of the baseline. Therefore, we believe that this study can help other future work to achieve more accurate stress detection results and thus help more people to solve stress-related problems.

Moreover, this work can contribute to other sensor data-based recognition tasks in the IMWUT community. Ubiquitous computing is a promising topic that involves sensors that passively measure data. Many of the applications can meet the same challenge in the limited number of labels compared to passive sensor samples. Thus, we believe that our work has the potential to be extended to multiple applications in this field.

7.7 Limitations

In this work, we have implemented several algorithms including the baselines, reproduced SOTAs, and the proposed semi-supervised methods. Although we tried to tune the model structure and hyperparameters for the evaluation datasets accordingly, however, there might still be room for improvement regarding the model performances. In addition, interpretability and explainability can be important when considering deploying the model into clinical applications. We have not explored interoperability at the current stage, which can be considered another limitation and future work.

8 CONCLUSION

In this work, we examined the SOTA semi-supervised algorithms and proposed a semi-supervised learning framework to help human stress estimation by leveraging massive unlabeled physiological and behavioral data collected in the wild. Our proposed method contained components of active sampling, auto-encoder pre-training, and consistency regularization. We evaluated our proposed methods using two datasets with a small amount of labeled data but a massive amount of unlabeled samples. We demonstrated that our proposed active sampling approach helped boost the performance of stress detection with fewer unlabeled samples. Our results showed that fully leveraging our proposed methods provided the best results in accuracy rate and F1 scores on both datasets. In addition, our ablation experiments on four human activity recognition datasets demonstrated that the proposed method improved the model performance compared to the supervised learning approach substantially. In the future, we will continue developing our method to improve its performance on datasets with categorical features and improve the interpretability and explainability of the methods.

ACKNOWLEDGMENTS

This work was supported by NSF (#1840167 and #2047296). We also thank the many researchers for sharing their datasets.

REFERENCES

[1] Kirstin Aschbacher, Aoife O'Donovan, Owen M Wolkowitz, Firdaus S Dhabhar, Yali Su, and Elissa Epel. 2013. Good stress, bad stress and oxidative stress: insights from anticipatory cortisol reactivity. *Psychoneuroendocrinology* 38, 9 (2013), 1698–1708.

[2] Dmitrijs Balabka. 2019. Semi-supervised learning for human activity recognition using adversarial autoencoders. In *Adjunct proceedings of the 2019 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2019 ACM international symposium on wearable computers*. 685–688.

[3] Brandon Ballinger, Johnson Hsieh, Avesh Singh, Nimit Sohoni, Jack Wang, Geoffrey H Tison, Gregory M Marcus, Jose M Sanchez, Carol Maguire, Jeffrey E Olgin, et al. 2018. DeepHeart: semi-supervised sequence learning for cardiovascular risk prediction. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[4] Rummana Bari, Md. Mahbubur Rahman, Nazir Saleheen, Megan Battles Parsons, Eugene H. Buder, and Santosh Kumar. 2020. Automated Detection of Stressful Conversations Using Wearable Physiological and Inertial Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 117 (dec 2020), 23 pages. https://doi.org/10.1145/3432210

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems* 32 (2019).

[6] Christopher M Bishop. 2006. Pattern recognition. *Machine learning* 128, 9 (2006).

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

[8] Sherilynn F. Chan and Annette M. La Greca. 2013. *Perceived Stress Scale (PSS)*. Springer New York, New York, NY, 1454–1455. https://doi.org/10.1007/978-1-4419-1005-9_773

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[10] Matteo Ciman and Katarzyna Wac. 2016. Individuals' stress assessment using human-smartphone interaction analysis. *IEEE Transactions on Affective Computing* 9, 1 (2016), 51–65.

[11] Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in neural information processing systems*. 3079–3087.

[12] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn Schuller. 2017. Semisupervised autoencoders for speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 1 (2017), 31–43.

[13] Firdaus S Dhabhar. 2014. Effects of stress on immune function: the good, the bad, and the beautiful. *Immunologic research* 58, 2-3 (2014), 193–210.

[14] Amr Gaballah, Abhishek Tiwari, Shrikanth Narayanan, and Tiago H Falk. 2021. Context-Aware Speech Stress Detection in Hospital Workers Using Bi-LSTM Classifiers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8348–8352.

[15] Shruti Gedam and Sanchita Paul. 2021. A review on mental stress detection using wearable sensors and machine learning techniques. *IEEE Access* 9 (2021), 84045–84066.

[16] Martin Gjoreski, Mitja Luštrek, Matjaž Gams, and Hristijan Gjoreski. 2017. Monitoring stress with a wrist device using context. *Journal of biomedical informatics* 73 (2017), 159–170.

[17] Jennifer A Healey and Rosalind W Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems* 6, 2 (2005), 156–166.

[18] Lee B Hinkle, Kamrad Khoshhal Roudposhti, and Vangelis Metsis. 2019. Physiological measurement for emotion recognition in virtual reality. In *2019 2nd International Conference on Data Intelligence and Security (ICDIS)*. IEEE, 136–143.

[19] Thomas H Holmes and Richard H Rahe. 1967. The social readjustment rating scale. *Journal of psychosomatic research* (1967).

[20] MaryCarol R Hunter, Brenda W Gillespie, and Sophie Yu-Pu Chen. 2019. Urban nature experiences reduce stress in the context of daily life based on salivary biomarkers. *Frontiers in psychology* 10 (2019), 722.

[21] Yash Jain, Chi Ian Tang, Chulhong Min, Fahim Kawsar, and Akhil Mathur. 2022. ColloSSL: Collaborative Self-Supervised Learning for Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 17 (mar 2022), 28 pages. https://doi.org/10.1145/3517246

[22] Kazuomi Kario, S McEWEN Bruce, and G PICKERING Thomas. 2003. Disasters and the heart: a review of the effects of earthquake-induced stress on cardiovascular disease. *Hypertension Research* 26, 5 (2003), 355–367.

[23] David N Khansari, Anthony J Murgo, and Robert E Faith. 1990. Effects of stress on the immune system. *Immunology today* 11 (1990), 170–175.

[24] Zachary D King, Judith Moskowitz, Begum Egilmez, Shibo Zhang, Lida Zhang, Michael Bass, John Rogers, Roozbeh Ghaffari, Laurie Wakschlag, and Nabil Alshurafa. 2019. micro-Stress EMA: A Passive Sensing Framework for Detecting in-the-wild Stress in Pregnant Mothers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 91.

[25] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. 2011. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter* 12, 2 (2011), 74–82.

[26] Samuli Laine and Timo Aila. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242* (2016).

[27] Boning Li and Akane Sano. 2020. Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–26.
[28] Dongxin Liu and Tarek Abdelzaher. 2021. Semi-supervised contrastive learning for human activity recognition. In *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 45–53.
[29] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods* 53, 4 (feb 2021), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y
[30] Mohammad Malekzadeh, Richard G Clegg, Andrea Cavallaro, and Hamed Haddadi. 2018. Protecting sensory data against sensitive inferences. In *Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems*. 1–6.
[31] Alban Maxhuni, Pablo Hernandez-Leal, L Enrique Sucar, Venet Osmani, Eduardo F Morales, and Oscar Mayora. 2016. Stress modelling and prediction in presence of scarce data. *Journal of biomedical informatics* 63 (2016), 344–356.
[32] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* 41, 8 (2018), 1979–1993.
[33] Karel Mundnich, Brandon M Booth, Michelle l'Hommedieu, Tiantian Feng, Benjamin Girault, Justin L'hommedieu, Mackenzie Wildman, Sophia Skaaden, Amrutha Nadarajan, Jennifer L Villatte, et al. 2020. TILES-2018, a longitudinal physiologic and behavioral data set of hospital workers. *Scientific Data* 7, 1 (2020), 1–26.
[34] Arthur Pimentel, Abhishek Tiwari, and Tiago H Falk. 2021. Human Mental State Monitoring in the Wild: Are We Better Off with DeeperNeural Networks or Improved Input Features? *CMBES Proceedings* 44 (2021).
[35] Beanbonyka Rim, Nak-Jun Sung, Sedong Min, and Min Hong. 2020. Deep learning in physiological signal data: A survey. *Sensors* 20, 4 (2020), 969.
[36] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-Task Self-Supervised Learning for Human Activity Detection. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 61 (jun 2019), 30 pages. https://doi.org/10.1145/3328932
[37] Wendy Sanchez, Alicia Martinez, and Miguel Gonzalez. 2017. Towards job stress recognition based on behavior and physiological features. In *International conference on ubiquitous computing and ambient intelligence*. Springer, 311–322.
[38] Pritam Sarkar and Ali Etemad. 2020. Self-supervised ECG representation learning for emotion recognition. *IEEE Transactions on Affective Computing* (2020).
[39] Yuan Shi, Minh Hoai Nguyen, Patrick Blitz, Brian French, Scott P. Fisk, Fernando De la Torre, Asim Smailagic, Daniel P. Siewiorek, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, and Santosh Kumar. 2010. Personalized Stress Detection from Physiological Measurements.
[40] Elena Smets. 2018. Towards large-scale physiological stress detection in an ambulant environment. (2018).
[41] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems* 33 (2020), 596–608.
[42] Se-Hui Song and Dong Keun Kim. 2017. Development of a stress classification model using deep belief networks for stress monitoring. *Healthcare informatics research* 23, 4 (2017), 285–292.
[43] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*. 127–140.
[44] Setareh Rahimi Taghanaki, Michael Rainbow, and Ali Etemad. 2022. Self-Supervised Human Activity Recognition with Localized Time-Frequency Contrastive Representation Learning. *arXiv preprint arXiv:2209.00990* (2022).
[45] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, Soren Brage, Nick Wareham, and Cecilia Mascolo. 2021. SelfHAR: Improving Human Activity Recognition through Self-Training with Unlabeled Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 36 (mar 2021), 30 pages. https://doi.org/10.1145/3448112
[46] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems* 30 (2017).
[47] Terry T Um, Franz MJ Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. 2017. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 216–220.
[48] Conny MA van Ravenswaaij-Arts, Louis AA Kollee, Jeroen CW Hopman, Gerard BA Stoelinga, and Herman P van Geijn. 1993. Heart rate variability. *Annals of internal medicine* 118, 6 (1993), 436–447.
[49] Juan Vazquez-Rodriguez, Grégoire Lefebvre, Julien Cumin, and James L Crowley. 2022. Transformer-Based Self-Supervised Learning for Emotion Recognition. *arXiv preprint arXiv:2204.05103* (2022).
[50] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. 2019. Interpolation Consistency Training for Semi-supervised Learning. (7 2019), 3635–3641. https://doi.org/10.24963/ijcai.2019/504

[51] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. (2019).

[52] Huiyuan Yang, Han Yu, and Akane Sano. 2022. Empirical Evaluation of Data Augmentations for Biobehavioral Time Series Data with Deep Learning. *arXiv preprint arXiv:2210.06701* (2022).

[53] Huiyuan Yang, Han Yu, Kusha Sridhar, Thomas Vaessen, Inez Myin-Germeys, and Akane Sano. 2022. More to Less (M2L): Enhanced Health Recognition in the Wild with Reduced Modality of Wearable Sensors. *arXiv preprint arXiv:2202.08267* (2022).

[54] Kangning Yang, Chaofan Wang, Yue Gu, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. 2021. Behavioral and Physiological Signals-Based Deep Multimodal Approach for Mobile Emotion Recognition. *IEEE Transactions on Affective Computing* (2021).

[55] Han Yu, Thomas Vaessen, Inez Myin-Germeys, and Akane Sano. 2021. Modality Fusion Network and Personalized Attention in Momentary Stress Detection in the Wild. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.

[56] Hang Yuan, Shing Chan, Andrew P Creagh, Catherine Tong, David A Clifton, and Aiden Doherty. 2022. Self-supervised Learning for Human Activity Recognition Using 700,000 Person-days of Wearable Data. *arXiv preprint arXiv:2206.02909* (2022).

[57] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).