Group-specific models of healthcare workers' well-being using iterative participant clustering

Vinesh Ravuri
Texas A&M University
College Station, TX
vineshr@tamu.edu

Projna Paromita
Texas A&M University
College Station, TX
projna.paromita@tamu.edu

Karel Mundnich
University of Southern California
Los Angeles, CA
mundnich@usc.edu

Amrutha Nadarajan
University of Southern California
Los Angeles, CA
nadaraja@usc.edu

Brandon M. Booth
University of Southern California
Los Angeles, CA
bbooth@usc.edu

Shrikanth S. Narayanan
University of Southern California
Los Angeles, CA
shri@sipi.usc.edu

Theodora Chaspari
Texas A&M University
College Station, TX
chaspari@tamu.edu

***Abstract*—Healthcare workers often experience stress and burnout due to the demanding job responsibilities and long work hours. Ambulatory monitoring devices, such as wearable and environmental sensors, combined with machine learning algorithms can afford us a better understanding of the naturalistic onset and evolution of stress and emotional reactivity in real-life with valuable implications in behavioral interventions. However, the typically large degree of inter-subject variability, due to individual differences in responses and behaviors, makes it difficult for machine learning models to robustly learn behavioral signal patterns and adequately generalize to unseen individuals. In this study, we design group-specific models of well-being (i.e., stress, sleep, positive affect, negative affect) and contextual outcomes (i.e., type of activity) based on real-life multimodal longitudinal data collected in situ from healthcare workers in a hospital environment. Group-specific models are constructed by learning an initial model based on all individuals and subsequently refining the model for a specific group of participants. Participants are originally grouped based on the feature space constructed by the multimodal data, while the original grouping is iteratively refined using the learned multimodal representations of the group-specific models. The results from this study indicate that in the majority of cases the proposed group-specific models, learned through iterative participant clustering, outperform the baseline systems, which involve general models learned based on all participants, as well as group-specific models without iterative participant clustering. This study provides promising results for predicting psychological and behavioral factors that affect the well-being of healthcare workers and lays the foundation toward ambulatory real-life assessment and interventions.**

***Index Terms*—Well-being, mental health, healthcare workers, group-specific machine learning, clustering**

I. Introduction

Healthcare workers often work in highly stressful environments that involve long hours and require rapid decision-making. This high-stress setting can have personal, social, and financial implications in healthcare workers [1], as well as an impact on their productivity, physical and emotional well-being, and overall quality of life [2]. Ambulatory monitoring, which relies on miniature sensor equipment to record normal daily activities, has been heralded as a promising solution to overcome limitations of in-lab studies related to retrospective bias and controlled recording conditions, allowing the opportunity to capture the naturalistic onset and dynamic processes of events that contribute to an individual's stress and emotional degradation [3]. Ambulatory data can provide insights for accurately characterizing clinically-relevant events and designing personalized interventions that contribute to augmenting work performance, mitigating negative mental health outcomes, and improving the overall healthcare [4].

Despite the promise of ambulatory assessment, designing technologies that can accurately track psychological and behavioral outcomes comes with various challenges. First, there is less control in real-life compared to laboratory conditions, therefore the resulting ambulatory data might be confounded by various factors, such as participants' activities and social interactions and environmental conditions [5]. Second, sensor reliability, connectivity, and recording placement are important factors that can highly skew the resulting data distributions [6]. Third, individual differences related to demographic, cognitive, and psychological factors can result in distinct manifestations of the same behavioral outcome across individuals, therefore, in significantly variable data distributions [7].

This paper explores the use of group-specific models to address the inherently large inter-individual variability. Group-specific models are specifically created for groups of participants with similar multimodal representations. We propose an iterative clustering approach that iteratively clusters participants based on the representations learned by the group-specific models, potentially providing more accurate and generalizable results compared to a one-time clustering. Clustering is based on K-Means performed on the feature space in the first clustering iteration, and the learned representation space in the following iterations. We evaluate our approach by estimating behavioral and psychological outcomes related to contextual factors (e.g., activity type) and well-being (i.e., stress, sleep, positive/negative affect) using multimodal data from 154 healthcare workers working in a real-world hospital environment [8].

978-1-7281-8699-3/20/$31.00 ©2020 IEEE
DOI 10.1109/TransAI49837.2020.00026

II. Prior Work

Subject and group-specific models depict a great potential in modeling the inherent inter-individual variability and providing accurate estimates of human outcomes. Subject-specific models are trained or fine-tuned using data from each individual, therefore achieving high predictive ability [9]. Yet, this approach has several limitations, since subject-specific models require labelled data for each person and each type of behavior, even if the latter occurs on a low-frequency basis. For this reason, subject-specific models might now always be generalizable to unseen individuals and conditions. Group-specific models have been proposed as an alternative to subject-specific models. They rely on grouping participants with similar characteristic into clusters, and subsequently learning a model for each cluster of participants. Prior work on group-specific models has focused on detecting stress, pain detection, and medication intake. Xu *et al.* and Koldijk *et al.* performed stress detection by learning K separate models for each participant cluster [10], [11]. Bertsimas *et al.* proposed the use of separate learners for each group of participants, which were jointly trained based on a target optimization criterion [12]. Lane *et al.* proposed an ensemble learning method to jointly recover the samples that are most valuable to each group of participants for an outcome of interest [13]. Finally, Gupta *et al.* proposed a conflict detection model between romantic partners and employed pre-defined individual traits, such as relationship satisfaction and attachment/avoidance, in order to group participants into theoretically postulated groups, followed by hierarchical and adaptive learning of group-specific models [14].

The contributions of this paper are the following: (1) We introduce a novel iterative participant clustering approach to group participants based on representations learned by the group-specific machine learning models, followed by learning of the group-specific models based on this clustering; and (2) We study a unique population (i.e., healthcare workers) performing complex and highly demanding tasks, a likely challenging dataset due to confounders from variable job responsibilities and high-pace work conditions.

III. Data Description

Our data come from a combination of self-reported, audio, and physiological data collected by 154 hospital workers over the course of ten weeks [8]. Participants engaged in their normal day-to-day activities, and data was collected through Internet-of-Things Bluetooth data hubs, wearable sensors, including a Fitbit Charge 2, OMsignal garment, and Unihertz Jelly Pro smartphone, as well as Ecological Momentary Assessments (EMAs), which were used to assess one's well-being, physical and psychological state, as well as contextual factors. Throughout the study, work-related questions were asked 31 times (every two days) and health-related questions were asked 35 times (every two days). Positive and negative affect were assessed using the ten items (five each) from PANAS-Short [15]. Stress was measured through a single question that read "Overall, how would you rate your current level of stress?" on a 1-5 scale. In order to obtain sleep information, participants were asked about the number hours they had slept the previous night. Activity was measured on a scale from 1-12 based on the amount of activity the participant was engaged in. Given the continuous nature of the outcomes of interest, the estimation of stress, sleep, positive/negative affect, and activity is treated as a regression problem in our study.

IV. Methodology

A. Data pre-processing

We used 25 activity-related features collected by the Fitbit Charge 2, 29 acoustic features from the Unihertz Jelly Pro, and 15 physiological features from the OMsignal garment [8], as summarized in Table I. All of the features were normalized using the min-max normalization.

B. Group-Specific Models

In the following, we will describe the proposed group-specific models, which rely on an iterative procedure between clustering participants into groups and fine-tuning a general model to each group. This process is summarized in Algorithm 1 and Figure 1.

1) Iterative Participant Clustering: The initial clustering is performed on the original 69-dimensional feature space. Participants were clustered using the mutual information between their sample distributions. Participants with similar values of mutual information were grouped in the same cluster. Following that and after the group-specific models finish the first round of training, we used K-means clustering based on the representation learned by the last hidden layer of each of the group-specific models. The similarity metric that was used in this clustering process relied on the Euclidean distance between the learned representations, which were averaged for each participant. This process was repeated ten times, and the highest correlation was reported.

2) Model Fine Tuning: Following each clustering iteration, we built group-specific models by first training a three-layer feedforward neural network (FNN) on all participants, and subsequently fine-tuning all layers of the model to each cluster of participants. Each iteration, the model goes through training and fine tuning. The model is fine tuned to learn weights more relevant to the group's specific training data while also preserving important representations from the entire data set. This process is done for all groups in the clustering, and repeated for each new cluster.

V. Experiments

A. Experimental Settings

The proposed method uses a three-layer FFN with 69 nodes in the first layer, 30 nodes in the second layer, and 1 node in the output. We use Early Stopping by monitoring the loss to determine when to stop training the model. The rectified linear activation function (ReLU) was used for all inner layers. The iterative clustering is repeated $N = 10$ times. Results are reported using $K = 3$ clusters, which was determined by the

TABLE I: Description of multimodal ambulatory features

Device	Feature description	# Features
Fitbit Charge 2	Upper/lower threshold of cardio activity range (CAR)*/fat burn activity range (FBAR)*/peak activity range (PAR)*/out of zone activity range (OORAR)*, number of minutes and calories burned in CAR/FBAR/PAR/OORAR, number of steps, minutes awake, minutes in deep/light/REM/non-REM sleep, minutes asleep, minutes in bed, sleep efficiency	25
OMsignal garment	Jitter, jitter 1st-order derivative, shimmer, fundamental frequency (average, average of smoothed contour, average of smoothed contour envelope), harmonic-to-noise ratio, voice probability, signal norm , signal norm computed with RelAtive SpecTrAl (RASTA)-Perceptual Linear Predictive (PLP) methodology, energy, zero-crossing rate, intensity, loudness, Fast Fourier transform (FFT) magnitude (250-650, 1000-4000Hz) sma, spectral roll-off of interquartile range (0-25%, 25-50%, 50-75%, 75-90%), FFT magnitude spectral flux, FFT magnitude spectral centroid/entropy/variance/skewness/kurtosis/slope, FFT magnitude sharpness/harmonicity	29
Unihertz Jelly Pro	Breathing rate, heart rate, intensity, average heart rate, average X/Y/Z acceleration, root mean square of first R-R interval difference, total power, very low/low/high frequency power, low to high frequency power	15

* OORAR/FBAR/CAR/PAR/: 0-50/50-69/70-85/85-100% of maximum heart rate

Algorithm 1: Iterative Clustering Algorithm

```
Initialize first cluster using on the original
  69-dimensional feature space (similarity metric:
  Mutual Information between participants' sample
  distributions)
N = Number of Iterations
K = Number of Groups
for i = 0 to N do
    for j = 0 to K do
        group data = corresponding j^th group data
        train model on all data
        fine tune model on group data
        if j == K − 1 then
            perform K-means on the learned data
              representations (similarity metric:
              Euclidean distance of samples averaged
              per participant)
        evaluate model
```

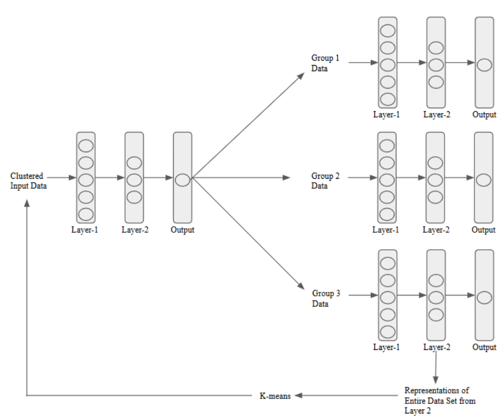

Fig. 1: Schematic representation of group-specific models with iterative participant clustering.

Dunn Index [16]. Evaluation is performed by comparing the actual and estimated values of the target outcomes through the Spearman's correlation coefficient.

TABLE II: Spearman's correlation coefficient between actual and estimated outcomes.

Construct	Baseline 1	Baseline 2	Baseline 3	Proposed
Stress	0.071	0.047	0.024	0.081
Sleep	0.363	0.402	0.320	0.372
Positive Affect	0.063	0.056	0.036	0.162
Negative Affect	0.029	0.015	-0.084	0.043
Activity	0.119	0.094	0.071	0.134

Differences between proposed and baseline are significant ($p \leq 0.01$).

B. Baseline Models

The proposed group-specific models with iterative participant clustering are compared to three baseline models. The first baseline is a general two-layer FFN with 69 nodes that is not fine-tuned on groups of participants (*Baseline 1*). The second baseline model is a three-layer FFN with 69 nodes in the first layer, 30 nodes in the second layer, and 1 node in the output layer, which employs the mutual information metric to perform an initial participant clustering, followed by a one-time fine tuning on each group of participants, but without iteratively clustering and re-learning on the new clusters (*Baseline 2*). The third baseline model is the same as the second baseline, but with the initial clustering determined by K-Means using a Euclidean distance between the mean of samples from each participant as a similarity metric (*Baseline 3*), rather than the proposed mutual information metric.

C. Results

Our results indicate that the proposed iterative participant clustering method yields better performance than the baselines for most constructs (Table II). The Activity outcome shows the least improvement possibly due to the fact that participant clusters did not change much after the first iteration. Negative affect showed improvement compared to the baseline approaches. Figure 2 shows the initial and intermediate participant clusters for the group-specific models of the negative affect outcome. Each instance also depicts the percent similarity between the previous and current cluster. We observe that although the negative affect construct depicts low performance, the iterative clustering procedure is able to learn representations for the outcome of interest that are more useful compared to the baseline methods. Positive affect depicts the greatest improvement with the proposed method having a difference in correlation of 0.099 from the best baseline.

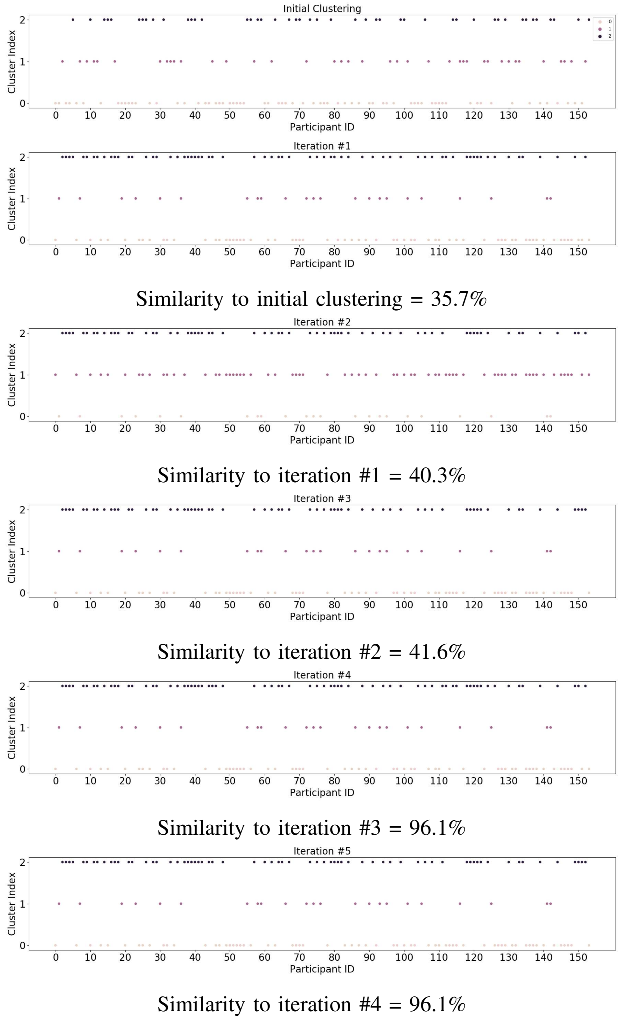

Fig. 2: Iterations of participant clustering for group-specific models of negative affect.

VI. Conclusion

We designed group-specific models of human outcomes for addressing the high inter-individual variability in ambulatory behavioral data. The proposed method that included an iterative participant clustering outperformed general models, as well as group-specific models that relied on a one-time participant clustering. Results provide a foundation for better understanding how machine learning combined with ambulatory data can help track healthcare workers' well-being.

Acknowledgement

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2017-17042800005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] A. Koinis, V. Giannou, V. Drantaki, S. Angelaina, E. Stratou, and M. Saridi, "The impact of healthcare workers job environment on their mental-emotional health. coping strategies: the case of a local general hospital," *Health psychology research*, vol. 3, no. 1, 2015.

[2] N. Embriaco, L. Papazian, N. Kentish-Barnes, F. Pochard, and E. Azoulay, "Burnout syndrome among critical care healthcare workers," *Current opinion in critical care*, vol. 13, no. 5, pp. 482–488, 2007.

[3] T. J. Trull and U. Ebner-Priemer, "The role of ambulatory assessment in psychological science," *Current Directions in Psychological Science*, vol. 23, no. 6, pp. 466–470, 2014.

[4] J. Shamian and F. El-Jardali, "Healthy workplaces for health workers in canada: knowledge transfer and uptake in policy and practice," *Healthcare Papers*, vol. 7, no. I, p. 6, 2007.

[5] M. Jones and D. Johnston, "Understanding phenomena in the real world: the case for real time data collection in health services research," *Journal of health services research & policy*, vol. 16, no. 3, pp. 172–176, 2011.

[6] R. W. Carpenter, A. M. Wycoff, and T. J. Trull, "Ambulatory assessment: New adventures in characterizing dynamic processes," *Assessment*, vol. 23, no. 4, pp. 414–424, 2016.

[7] G. Quer, P. Gouda, M. Galarnyk, E. J. Topol, and S. R. Steinhubl, "Inter-and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, bmi, and time of year: Retrospective, longitudinal cohort study of 92,457 adults," *Plos one*, vol. 15, no. 2, p. e0227709, 2020.

[8] K. Mundnich, B. M. Booth, M. l'Hommedieu, T. Feng, B. Girault, J. L'Hommedieu, M. Wildman, S. Skaaden, A. Nadarajan, J. L. Villatte *et al.*, "Tiles-2018: A longitudinal physiologic and behavioral data set of hospital workers," *arXiv preprint arXiv:2003.08474*, 2020.

[9] T. Chaspari, A. C. Timmons, and G. Margolin, *Population-Specific and Personalized (PSP) Models of Human Behavior for Leveraging Smart and Connected Data.* Taylor & Francis, 2019.

[10] Q. Xu, T. L. Nwe, and C. Guan, "Cluster-based analysis for personalized stress evaluation using physiological signals," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 275–281, 2014.

[11] S. Koldijk, M. A. Neerincx, and W. Kraaij, "Detecting work stress in offices by combining unobtrusive sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 227–239, 2016.

[12] D. Bertsimas, N. Kallus, A. M. Weinstein, and Y. D. Zhuo, "Personalized diabetes management using electronic medical records," *Diabetes care*, vol. 40, no. 2, pp. 210–217, 2017.

[13] N. D. Lane, Y. Xu, H. Lu, S. Hu, T. Choudhury, A. T. Campbell, and F. Zhao, "Community similarity networks," *Personal and ubiquitous computing*, vol. 18, no. 2, pp. 355–368, 2014.

[14] K. Gupta, A. Gujral, T. Chaspari, A. C. Timmons, S. Han, Y. Kim, S. Barrett, S. Sichko, and G. Margolin, "Sub-population specific models of couples' conflict," *ACM Transactions on Internet Technology (TOIT)*, vol. 20, no. 2, pp. 1–20, 2020.

[15] A. Mackinnon, A. F. Jorm, H. Christensen, A. E. Korten, P. A. Jacomb, and B. Rodgers, "A short form of the positive and negative affect schedule: Evaluation of factorial validity and invariance across demographic variables in a community sample," *Personality and Individual differences*, vol. 27, no. 3, pp. 405–416, 1999.

[16] C. Legány, S. Juhász, and A. Babos, "Cluster validity measurement techniques," in *Proceedings of the 5th WSEAS international conference on artificial intelligence, knowledge engineering and data bases*. World Scientific and Engineering Academy and Society (WSEAS), 2006, pp. 388–393.