frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 09 June 2023
DOI 10.3389/fdgth.2023.1195795

OPEN ACCESS

EDITED BY
Heysem Kaya,
Utrecht University, Netherlands

REVIEWED BY
Alena Velichko,
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Russia
Hatice Kose,
Istanbul Technical University, Türkiye

*CORRESPONDENCE
Projna Paromita
projna.paromita@tamu.edu

RECEIVED 28 March 2023
ACCEPTED 22 May 2023
PUBLISHED 09 June 2023

CITATION
Paromita P, Mundnich K, Nadarajan A, Booth BM, Narayanan SS and Chaspari T (2023) Modeling inter-individual differences in ambulatory-based multimodal signals via metric learning: a case study of personalized well-being estimation of healthcare workers. Front. Digit. Health 5:1195795. doi: 10.3389/fdgth.2023.1195795

Modeling inter-individual differences in ambulatory-based multimodal signals via metric learning: a case study of personalized well-being estimation of healthcare workers

Projna Paromita[1]*, Karel Mundnich[2], Amrutha Nadarajan[2], Brandon M. Booth[2], Shrikanth S. Narayanan[2] and Theodora Chaspari[1]

[1]HUman Bio-Behavioral Signals Lab, Texas A & M University, College Station, TX, United States, [2]Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, CA, United States

Introduction: Intelligent ambulatory tracking can assist in the automatic detection of psychological and emotional states relevant to the mental health changes of professionals with high-stakes job responsibilities, such as healthcare workers. However, well-known differences in the variability of ambulatory data across individuals challenge many existing automated approaches seeking to learn a generalizable means of well-being estimation. This paper proposes a novel metric learning technique that improves the accuracy and generalizability of automated well-being estimation by reducing inter-individual variability while preserving the variability pertaining to the behavioral construct.
Methods: The metric learning technique implemented in this paper entails learning a transformed multimodal feature space from pairwise similarity information between (dis)similar samples per participant via a Siamese neural network. Improved accuracy via personalization is further achieved by considering the trait characteristics of each individual as additional input to the metric learning models, as well as individual trait base cluster criteria to group participants followed by training a metric learning model for each group.
Results: The outcomes of the proposed models demonstrate significant improvement over the other inter-individual variability reduction and deep neural baseline methods for stress, anxiety, positive affect, and negative affect.
Discussion: This study lays the foundation for accurate estimation of psychological and emotional states in realistic and ambulatory environments leading to early diagnosis of mental health changes and enabling just-in-time adaptive interventions.

KEYWORDS
metric learning, Siamese neural network, healthcare workers, well-being, mental health, ambulatory monitoring

1. Introduction

Healthcare workers often experience significant job strain as a result of high physical, mental, and emotional workloads with low decision latitude and extensive responsibilities (1). These job characteristics can lead to high burnout rates, which can be the source of stress, anxiety, and depression (2). In light of the 2019 novel coronavirus (SARS-CoV-2) pandemic,

commonly known as the COVID-19 pandemic, the importance of the well-being of healthcare workers has become a primary concern (3). There has been a wide array of interventions at the institutional and personal level attempting to combat burnout of healthcare workers (4). Many medical schools and hospitals have shifted their focus on institutional policies that promote healthy work conditions, including hiring additional staff for administrative duties, introducing counseling services, mental health awareness campaigns, mentoring meetings, and group discussions with faculty and staff (5–8). Apart from these, mindfulness training and social cohesion wellness programs (e.g., mandatory wellness retreats) have been particularly important in promoting personal and social well-being (9–14). Such initiatives can create safe spaces for healthcare workers to share their experiences and help hospital workers become more aware of their thoughts and feelings, thus effectively reducing the adverse effects of burnout and promoting a healthy and supportive work environment. Yet, there is no "one-size-fits-all" approach to these interventions, which rather need to be tailored to the population of interest and its corresponding capabilities and characteristics. Moreover, attending the aforementioned programs and training sessions after long work hours often becomes burdensome for healthcare workers, rendering these initiatives ineffective in many cases.

Just-in-time adaptive interventions (JITAIs) are an emerging new intervention design for mental health that aims to provide the right type and amount of support at the right time, depending on a user's emotional state and other contextual factors (15–17). JITAIs leverage ambulatory devices, such as smartphones and wearable sensors, offering the means to track the moods and emotions of healthcare professionals in a personalized manner. They can assess one's behavior in a manner that is often inaccessible in standard clinical settings, providing ecologically valid data of complex cognitive, affective, and psychological processes over time for a particular individual, and potentially contributing to the early detection of health degradation (18, 19). Ambulatory data can provide a foundation for personalized interventions for healthcare workers that can potentially mitigate negative mental health outcomes, augment their work performance, and improve the overall care received by the patients in the hospital (20). Ambulatory data can further offer an alternative to self-reports (21), which are prone to response bias and recall bias, often resulting in under-reporting or over-reporting symptoms (22). Previous studies indicate that with proper prior testing and troubleshooting of the deployed ambulatory systems, healthcare workers overall respond positively to utilizing ambulatory sensors for monitoring wellness (23).

Several challenges affect ambulatory monitoring in the wild. Ambulatory data are collected in uncontrolled real-world environments and might be affected by various confounding factors, such as co-occurring activities (e.g., exercise), interpersonal relationships (e.g., conflict with co-workers), and environmental conditions (e.g., indoor ventilation, outdoor weather conditions) (24). The signals recorded by ambulatory sensors are further prone to non-standardized factors in the data collection process, including sensor misplacement and non-fixed physiological baseline (25), which result in highly variable data distributions (26). Finally, inherent individual differences (e.g., in terms of demography and psychological factors) often influence the collected human behavioral data and can be manifested across various signal modalities (27, 28). These impose significant challenges to the design of machine learning methodologies that can reliably quantify facets of psychological and emotional outcomes from ambulatory data.

We may address the inherent individual differences in signals collected through ambulatory sensors by constructing models that reduce the evidence of person-specific information on the data while preserving information about the behavioral outcome of interest (29). Toward this end, personalized and group-specific models have recently been the focus of multiple research efforts, since they provide tailored estimates of behavioral constructs for each participant or group of participants. Personalized models are tailored to a specific participant via learning their individual behavioral patterns (30), as well as integrating their demographic and anthropomorphic characteristics (31, 32). Such approaches usually involve the learning of separate models for each participant (33), which entails the risk of over-fitting or inadequate training due to the lack of enough data from a given subject. Group-specific models cluster participants according to clinically and theoretically relevant characteristics and are usually implemented through hierarchical and adaptive learning methods (34–36). Such methods use a portion of data from a target entity to fine-tune the decisions of a machine learning model adaptively (37, 38). Increasing the amount of target data used for fine-tuning the models tends to improve their performance (39). Metric learning techniques, such as Siamese neural networks (SNN), require a small amount of data to fine-tune such models as they model the relative distance between samples, rather than the absolute class-wise patterns (40). Thus, metric learning methods are a promising solution to personalized and group-specific models when the labeled samples from a target entity are scarce.

In this study, we propose a metric learning technique for estimating the well-being of healthcare workers in a personalized manner. Metric learning is implemented with an SNN architecture that learns a multimodal signal transformation that models the pairwise similarity between similar and dissimilar samples from each participant. The proposed pairwise similarity approach does not require the separate estimation of the data distribution for each class, therefore it can be more robust to outliers and more forgiving to small datasets (41). We have further examined the effect of including an increasing number of labeled samples from the target participants to fine-tune the SNN. Finally, as an additional personalization step, we have integrated participants' trait characteristics in two ways: (1) as an input to the models; and (2) as a clustering criterion to group participants, followed by training separate models for each group. Our proposed approach is evaluated on a dataset collected in a real-world hospital environment over time, including data from 139 healthcare workers with constructs related to well-being (i.e., anxiety, stress, positive/negative affect) (42). Our findings suggest that the proposed approach preserves the variability attributed to the behavioral outcome of interest yielding promising results in terms of estimating behavioral constructs while minimizing user-

dependent variability. The proposed metric learning approach further outperforms non-personalized models, such as a feedforward neural network (FNN) trained on all participants using the original ambulatory-based features, as well as conventional distribution-based adaptation methods, such as the maximum independence domain adaptation (MIDA), that attempts to mitigate inter-individual differences by aligning feature distributions from all participants. Implications of this study result in the accurate longitudinal tracking of well-being in highly demanding job settings and help lay a foundation to apply JITAIs for mental health (17).

2. Prior work

Ambulatory devices can unobtrusively collect data from participants in their daily lives, providing us with informative insights regarding daily activities, physical outcomes, and behavior patterns, therefore informing treatment and intervention procedures (43, 44). Smartphones are widely used devices that provide an effective way to quantify behavior via the various in-built sensors, while at the same time integrating the computational power to analyze the signals captured by the sensors in a meaningful manner (45). In a study conducted by Hunasgi et al. (46) aiming to recognize stress in dental students, an Android smartphone app called "S-HEALTH" has been used to detect heart rate, oxygen saturation, and stress levels via smartphone sensors. The authors collected ambulatory-based data before and after a stressful event, during which physiological recordings have been collected by asking participants to touch a specific sensor of the smartphone device. Apart from collecting physiological data, smartphones have been also used to collect other behavioral metrics, such as user feedback data, overall phone activity, speech, location, and physical movement, which play a key part in detecting stress (47–49), anxiety (50), negative affect (51), unusual activities (52), and mood (53). Wearable devices can further enable physiological recording in a real-world setting (i.e., "in-the-wild"). Data collected through activity trackers, such as Fitbit (54), have been used to predict health and wellbeing status (55), detect daily stressful events (56) or predict future mood (57–59). Schmid et al. have used wearable sensors to track healthcare workers' heart rate variability (HRV) during mindfulness exercises (60). Sano et al. have used wrist-based sensors to record the physical activity and autonomic physiology of college students over the span of a month with the goal of assessing stress and mental health (61). Smartphones in combination with wearable devices have been used in several other studies to detect stress and depression (47, 62, 63). Gaballah et al. proposed a context-aware speech stress detection model using the TILES-2018 dataset (42), which is the same dataset we have used in our study. Gaballah et al.'s study utilizes audio, location, and circadian rhythm signals extracted from different ambulatory devices, for example, smartphones, smartwatches, and smart garments, from 144 healthcare workers working for 10 weeks to detect stress in the wild. Their proposed bidirectional long short-term memory (LSTM) model reached a classification F1 score of up to 65%, while environmental sensor data acted as the context for the model (64). Another study uses physiological, physical, sleep, and context features from TILES-2018 and the TILES-2019 datasets (65) to develop personalized models to predict future affect. They achieve personalization by developing individual-specific random forest and mixed-effect random forest models. These models were trained using the data collected per individual during the first 2–8 weeks of the total 10-week data collection period and predicted future affect reaching correlation coefficients up to 0.8 for positive affect and 0.5 for negative affect (66). Multi-modal signals collected using an assortment of ambulatory devices from a large ($N = 606$) number of individuals, were used to detect stress in the wild in a separate study. Through a regression analysis, this study reached a Spearman correlation of 0.25 using a combination of random forest and state-trait anxiety inventory to detect stress (67).

The data collected through ambulatory devices often demonstrates high-inter individual variability, which can be addressed via the design of personalized models that can accurately model physical and behavioral outcomes while at the same time eliminating the evidence of inter-individual variability in the data attributed to behaviorally non-relevant information (51, 68, 69). Personalized machine learning models are popular in various domains, such as in computer vision where models have been used to capture the inherent individual diversity in detecting facial expressions and gestures from unlabeled data (37) or to generate automatic tags for a video based on a user's comments and preferences (70). Personalized models implemented with multitask learning have been also employed to recognize human activity from multimodal sensor data (71, 72) and in the detection of unusual events pertaining to a user's activity using speech data recorded via smartphone devices (52). Similarly, De Santos et al. have utilized personalized models to increase bio-metric system security via detecting deviation from normal behavior for accurate person identification (73). While studying the impact of meal consumption on one's glucose response, Zeevi et al. have opted to use personalized models given the high inter-individual variability of the glucose response signal (69). Personalized models have been also employed to develop treatment regimens, where such models have been trained to optimize a target criterion, such as maintaining blood glucose at a healthy level (74, 75). They have been further utilized to detect momentary negative affect of undergraduate students via incorporating contextual data (e.g., location, movement, phone usage) collected from smartphone devices (51), and have been found to outperform generalized models for the same task (58, 76).

Personalized models may be developed following various procedures, for example, via modeling data separately for each individual (30), via transfer learning (by fine-tuning a general model using data from a target individual) (37), or via integrating individual factors to the model input, such as anthropomorphic, psychological, and cognitive characteristics (32, 69). Psychological studies have found several individual factors, such as coping styles and current life conditions (e.g., poor family relationships), to be related to stress moderation, and physiological and affective responses (77–79), motivating the

integration of such factors into the design of personalized machine learning models. A relatively small number of studies have considered individuals' trait characteristics when modeling behavioral outcomes. Gupta et al. and Yadav et al. have leveraged individuals' trait characteristics, such as trait anxiety, personality, and attachment variables, to form clusters of participants, based on which they have designed population-specific models of behavior (36, 80). In supporting sleep health, Nguyen et al. have utilized personality traits and chronotype characteristics (i.e., one's morning or evening preference) to develop personalized feedback for sleep interventions (81). Tondello et al. have further used five gaming traits that reflect participants' gaming preferences in developing a personalized gaming platform (82). Individual characteristics, such as ideology, dominant emotional sentiments, and personality have been further utilized to develop a personalized psychological intervention system to promote inter-group relations (83). This evidence suggests that considering individuals' characteristics is an important design factor toward personalized behavior detection models and subsequently, effective JITAIs.

Data collected from ambulatory devices are prone to errors related to ambient noise and sensor misplacement. In addition, it is usually difficult to obtain labels from all samples during ambulatory monitoring studies, therefore labeled data are not always sufficient for modeling absolute patterns in data distributions for each participant (24). Few-shot learning techniques that are commonly based on metric-learning approaches model the relative distance between samples, thus can be trained with a small amount of data (40). These models have the ability to naturally rank data samples based on sample pair similarity and have shown promising results in image recognition (84), visual tracking (85), and detection of the continuous spectrum of disease severity (86). Siamese neural networks (SNNs) have also found applications in speech-based emotion recognition (87, 88) and gait-based user identification (89). Although SNNs have been employed extensively for classification purposes, regression analysis using SNNs is a field that is not extensively studied (90). Moreover, none of the previous works have formulated such techniques in the context of personalized learning of behavioral outcomes.

The contributions of this paper are the following: (1) we introduce a novel metric learning approach implemented with an SNN regressor to learn personalized representations of multimodal signals and their association with behavioral outcomes; (2) we investigate additional ways to personalize our models by utilizing the participants' trait characteristics and fine-tune the models using a portion of the data collected from the target participants; and (3) we evaluate these methods on ecologically valid data collected in situ from healthcare professionals with variable job responsibilities and high-pace work conditions.

3. Data description

Our data is part of the TILES-2018 dataset (42), which was collected from 212 full-time hospital workers, aged between 21–65 years, over 10 weeks. Participants engaged in their typical daily activities while being equipped with ambulatory wearable devices and sensors to collect vocal acoustic and physiological signals throughout the period of data collection. A Fitbit Charge 2 was used to measure sleep activity and exercise, an OMsignal garment collected heart rate and breathing rate, and the Unihertz Jelly Pro smartphone, a small and lightweight phone worn on the lapel, was programmed to obtain vocal acoustic features from statistically sampled egocentric audio recordings (91).

Participants completed several surveys during the enrollment and data collection period. Initial ground truth battery (IGTB) surveys were collected once during the enrollment period to assess participants' trait characteristics related to job performance, cognitive abilities, and health. Cognitive IGTB constructs include participants' organization citizenship behaviors (OCB), which were measured using the Organization Citizenship Behavior Checklist (OCB-C) (92) via 20 items on a scale ranging from 1–5, and cognitive ability, which was captured using both the Shipley Abstraction Test (93) assessing fluid intelligence (25 items on a scale of 0–25) and the Shipley Vocabulary test assessing crystallized intelligence. In the Shipley Vocabulary test, participants would match words with their synonyms from a list of 40 items, and the score is calculated from the correctness of the matching between the words. Psychological IGTB constructs include personality traits (i.e., extraversion, agreeableness, conscientiousness, emotional stability, openness), assessed via the Big Five Inventory-2 (BFI-2) (94), trait affect (i.e., the participants' overall affect that does not depend on momentary factors) captured with the Positive and Negative Affect Schedule Expanded form (PANAS-X) (95) using 60 items on a scale of 1–5, and trait anxiety measured by the Trait-Form of the State Trait Anxiety Inventory (STAI) (96). Apart from this, we have utilized the tobacco usage data, collected through the 3-item Global Adult Tobacco Survey (GATS) (97) and sleep quality data, collected through the 19-item Pittsburgh Sleep Quality Index (PSQI) (98) for our experiments. The IGTB measures are used as an additional input to the machine learning models, as well as a clustering criterion to group participants, followed by training separate models for each group.

Momentary ground truth (MGT) surveys of well-being constructs were collected once per day through a smartphone app using ecological momentary assessments (EMAs) related to participants' anxiety, stress, positive affect, and negative affect. To make the daily self-reporting more accessible to the participants, the number of assessment items is reduced while following the prior standard scaling format with significant reliability (42). Daily measurement of anxiety and stress was conducted by asking two questions (i.e., "Please select the response that shows how anxious you feel at the moment" and "Overall, how would you rate your current level of stress?") and were both assessed on a 5-point Likert scale. Affect was obtained through the Positive and Negative Affect Schedule-Short form (PANAS-S) (99), which includes 10 items, 5 items each for positive and negative affect. Each item of this questionnaire was scaled from 1 (very slightly or not at all) to 5 (extremely), resulting in a label value ranging between 5–25. This study uses the MGT constructs of anxiety,

stress, positive affect, and negative affect as the behavioral constructs of interest, that represent participants' well-being outcomes. The corresponding distributions of these constructs are presented in **Figure 1**, where each x-tick represents the discrete values of the construct (e.g., self-reported anxiety MGT is labeled between 1–5, thus there are 5 representations in **Figure 1A**) and y-tick represents the total count of some particular label present in the full dataset.

Much effort was spent before and during the data collection to maximize the quality and ecological validity of the data (100). Despite this effort, some of the participants had missing data from one or more sensors as well as missing ground truth information from the self-reports. For this study, 73 participants had less than 13 days of self-reported data out of the 10-week study period, where 13 is an empirically chosen minimum number of days for which participants require self-reports. Since self-reports act as a ground truth for the proposed machine learning methods and proper ground truth information is necessary for supervised learning, we removed these participants from the total pool, resulting in 139 participants with on average 3-week data that were included in the experiments. From these 139 participants, around 10% of feature values on average per participant were missing. These feature values served as the input to the machine learning models and were replaced using linear interpolation **Section 4.1**.

4. Methodology

This section delves into details of processing the data, designing the proposed models, and evaluating our approach. **Section 4.1** contains details on how we pre-processed our data and extracted features. The loss functions to achieve personalization through metric learning and the models developed for different experiments are documented in the **Section 4.2**. **Section 4.3** describes the details of our experimental framework and **Section 4.4** refers to the baselines used to compare the performance of our proposed models.

4.1. Data pre-processing and feature extraction

The input features of the different models included both ambulatory measures and participants' trait characteristics. We used 69 ambulatory features available as part of TILES-2018 dataset, which include 25 measures of physiology and daily activity from Fitbit Charge 2, 29 acoustic features from Unihertz Jelly Pro, and 15 physiological features from OMsignal garment (42) (**Table 1**). We accounted for the missing values at the sample level of the features by replacing the missing data points through linear interpolation using the data points before and after the missing

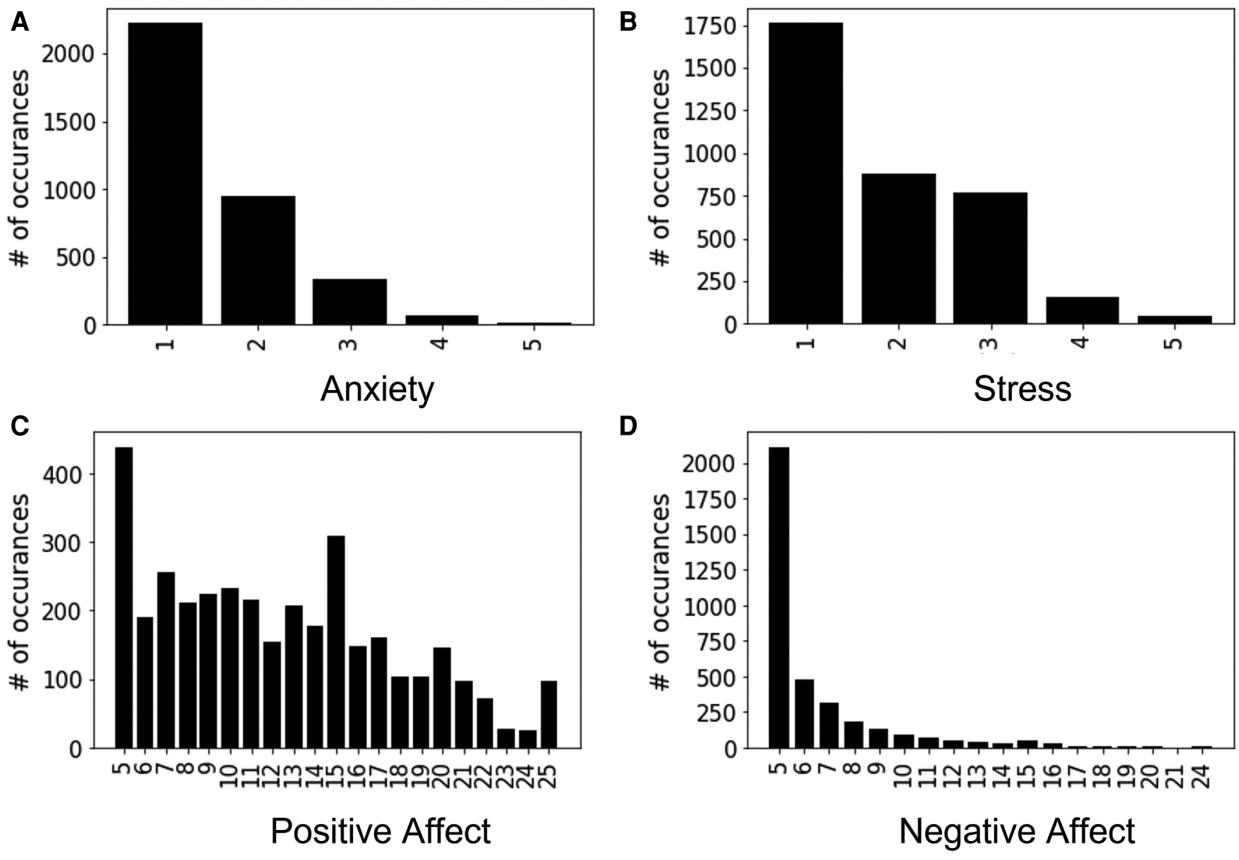


FIGURE 1
Distribution of momentary ground truth (MGT) data used as outcomes.

TABLE 1 Description of ambulatory features recorded continuously over 10 weeks via wearable devices.

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
Unihertz Jelly Pro	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, 1000–4000 Hz), 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
OMsignal garment	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.

data. We further calculated the daily average of these features for our analysis. Figure 2 depicts the distribution of two example features, namely, the resting heart rate and the minimum number of burned calories, for each participant. We observe large inter-individual differences in these features' range and distribution shape, providing evidence of the need to design personalized models.

Furthermore, as an additional input to the proposed personalized models, we have considered 14 IGTB features that are indicative of participants' trait characteristics, as described in Section 3 and summarized in Table 2. All the samples used in our experiments are independent over time. Also, we conducted min-max feature normalization to these ambulatory features using the minimum/maximum value of the corresponding training set's participants to ensure uniformity while using them for training our proposed models.

4.2. Metric learning for personalized models

The metric learning approach, which we implement via SNN, learns the relative distance between two items in a pair based on whether the items are similar or dissimilar. So, we form pairs of samples to train our proposed model. The pairs could be formed using two similar or dissimilar instances of input (x) as described in Section 4.2.1. One of the ways we achieve personalization in this study is by rigorously following a participant-wise format while forming the pairs. The metric learning is achieved via equation 1. Here we calculate the relative distance between the transformed embeddings ($\mathbf{f_W}(\mathbf{x}_{nm})$ and $\mathbf{f_W}(\mathbf{x}_{nm'})$) and the normalized ground truths ($\hat{y}_{nm}$ and $\hat{y}_{nm'}$) among the pairs we formed for our SNN structure. Subsequently, the transformed embedding ($\mathbf{f_W}(\mathbf{x}_{nm})$) is used in the regression analysis, where through another layer of transformation ($\mathbf{g_V}$), the learned embedding helps in predicting the constructs considered in this study. The regression loss, described in equation 2, helps in updating the whole structure for better performance. In the following, we will describe the metric learning formulation that has been used to implement the proposed personalized models.

Let $\mathcal{D}_n = \{\mathbf{x}_{nm}, y_{nm}\}$ be the set of data obtained from participant n, where $\mathbf{x}_{nm}$ and y_{nm} represent the feature vector and MGT label of the mth sample from participant n. Let $\hat{y}_{nm}$ represent the ℓ_2-normalized version of y_{nm}. Metric learning

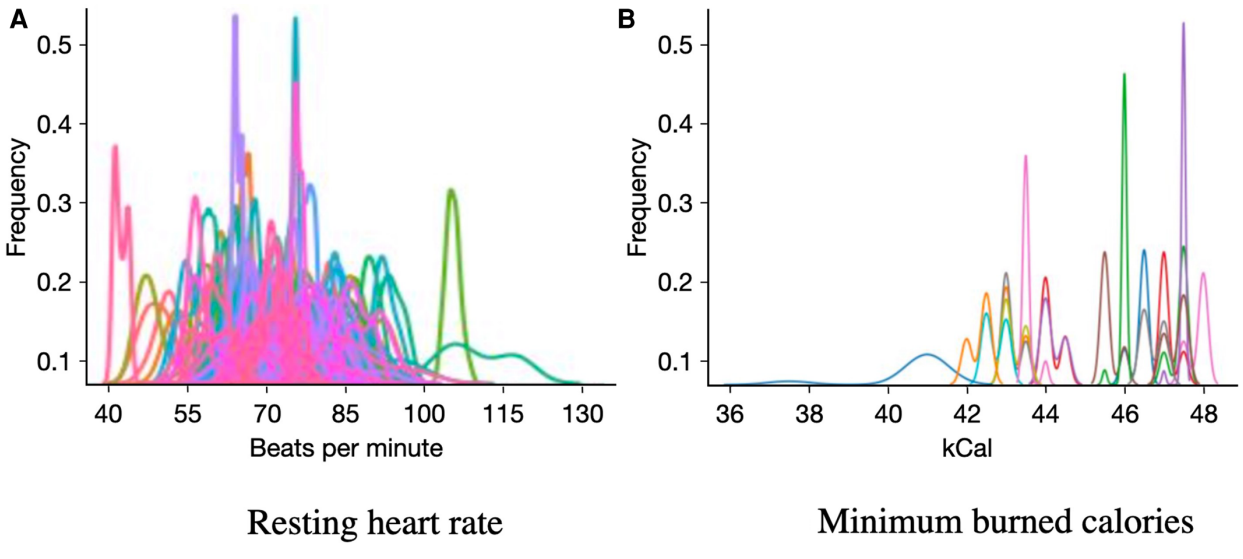


FIGURE 2
Histograms of feature distributions measured from the ambulatory data color-coded by participants.

TABLE 2 Description of initial ground truth battery (IGTB) features collected once during the enrollment period.

IGTB type	IGTB description	# features
Cognitive	Organization citizenship behaviors, fluid intelligence, crystallized intelligence	3
Psychological	Extraversion, agreeableness, conscientiousness, emotional stability, openness, trait positive affect, trait negative affect, trait anxiety	8
Health	Tobacco use (Yes/No), Tobacco quantity, sleep quality	3

learns a transformation (or embedding) $\mathbf{f_W}$ of the input feature, parameterized through $\mathbf{W}$, such that pairs of samples corresponding to proximal levels of the MGT outcomes are projected to the same neighborhood of the feature embedding, while the opposite occurs for samples with a distinctively large difference in the MGT outcome. This is implemented via the following loss function:

$$l_F = \sum_n \sum_{m \neq m'} |(\|\mathbf{f_W}(\mathbf{x}_{nm}) - \mathbf{f_W}(\mathbf{x}_{nm'})\|_2 - \|\hat{y}_{nm} - \hat{y}_{nm'}\|_2)| \quad (1)$$

where (1) defines the difference between the distance of samples m and m' of participant n with respect to the transformed feature space f_W and the MGT label space. The distance with respect to both the embedding and label space is measured via the Euclidean distance, where $\|\cdot\|_2$ represents the ℓ_2-norm.

In order to impose additional supervision and achieve a reliable estimation of the MGT outcomes, the feature embedding $\mathbf{f_W}$ is further transformed to the final MGT outcome via $\mathbf{g_V}$, parameterized through weights $\mathbf{V}$. The parameters included in $\mathbf{V}$ are learned using the following loss function, which minimizes the mean square error between the actual and estimated MGT output:

$$l_R = \sum_n \sum_m \|\mathbf{g_V}(\mathbf{f_W}(\mathbf{x_{nm}})) - y_{nm}\|_2^2 \quad (2)$$

The above optimization problem can be implemented via an SNN (Figure 3A). The SNN takes as an input pair of samples from the same participant and projects those into a feature embedding via the transformation $\mathbf{f_W}$ in the shared $\mathbf{f_W}$ layer of the schematic diagram (Figure 3), which minimizes the feature loss l_F. The output of the SNN is an MGT outcome resulting from transforming the feature embedding via the transformation $\mathbf{g_V}$, which minimizes the regression loss l_R.

Utilizing this formulation, we conducted an ablation study comprising three models: (1) metric learning with ambulatory features (ML-A): this model takes the ambulatory features as input to the model (Section 4.2.1), (2) metric learning with ambulatory and trait features (ML-AT): this model takes IGTB features representing participants' trait characteristics, along with the ambulatory features as input to the model (Section 4.2.1), and (3) separate metric learning models with ambulatory and trait features for each IGTB cluster of participants (ML-C): after developing clusters of data based on IGTB values, this model applies the ML-AT model on separate clusters (Section 4.2.2). A detailed description of each of these is provided below.

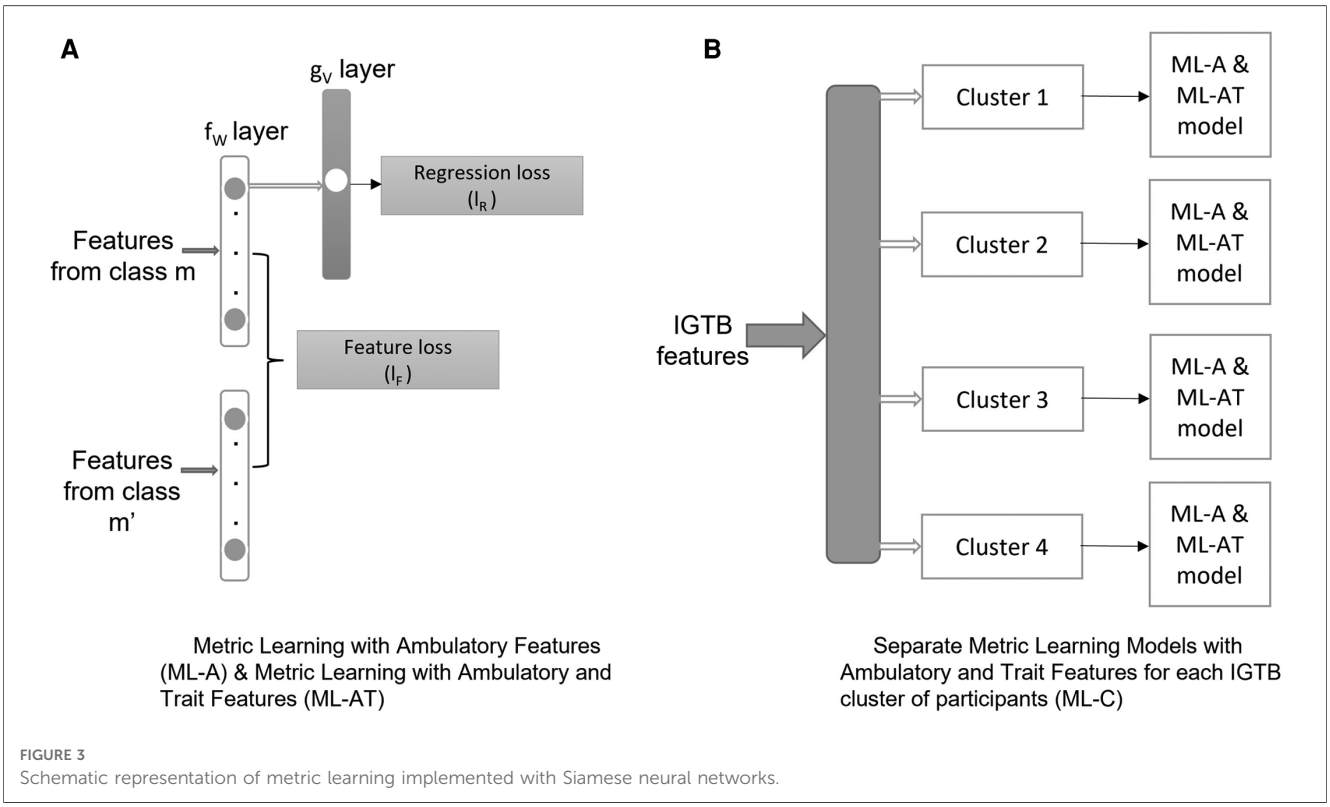


FIGURE 3
Schematic representation of metric learning implemented with Siamese neural networks.

4.2.1. Metric learning with ambulatory features (ML-A)

As the first model of our ablation study, ML-A incorporates personalization through personalized pair formation (Section 4.2) and transfer learning. We study this model to understand how personalized models using only the ambulatory features performs in predicting different job constructs. ML-A was implemented with an SNN whose input includes the 69 ambulatory features described in Table 1. For training the SNN, the original dataset was divided into 10 stratified folds with each fold containing around 14 participants. We conducted the "participant-independent cross-validation" process for training and testing our models, thus there was no participant overlap across folds. Among the folds we developed, one fold is utilized for validation, one fold for testing, and the remaining were used to train the model in a looping fashion ensuring we have tested all the folds using separate models. We ensured that no overlap exists among training, validation, and testing folds, thus allowing us to explore the generalizability of the models to unseen samples. Moreover, all the model parameters were removed after each loop for testing each of the folds to avoid data leakage.

In order to form the pairs to train the model, we defined two labels of the instances to form the pair as being similar if $|a - b| \leq 1\ \forall a, b \in$ *labels*, and the rest dissimilar given the condition is not met empirically. Although positive affect and negative affect have a considerably different range of labels than anxiety and stress, outcomes of experimentations, while changing the threshold for similarity from 1 to 5, did not show a significant difference in performance. Moreover, this has helped us tackle the problem of the four target constructs having different ranges of labels (anxiety: 1–5, stress: 1–5, positive affect 5–25, negative affect: 5–24).

To further personalize our model, we utilized the transfer learning process by taking 0–90% of testing data (i.e., including a percentage of samples between 0–90% from the target participants in the test fold), referred to as "fine-tuning data," and included that in training fold for the purpose of fine-tuning our model. This part was done after the hyper-parameters were set via the hyper-parameter tuning process (explained in Section 4.3) for the given fold and at the beginning of the SNN training. We removed the "fine-tuning data" from the target participants' data of the testing folds to ensure that any data leakage is completely avoided during the training and testing of the models. Algorithm 1 outlines the training process of the ML-A method. It depicts one single iteration of the algorithm which has the outer loop looping through F folds. Inside the outer loop, there are two separate loops, the first loop helps in preparing the pairs, and the second loop trains the model until the necessary condition for the early stopping is met. With random initialization, we run this algorithm 20 times for this model.

subsubsectionMetric learning with ambulatory and trait features (ML-AT)

In the second model, ML-AT, of our ablation study, we introduced personalizing our models through additional features as input to the model. These additional features are the trait characteristics (IGTB) (Table 2) features, which were included along with ambulatory features as an input to the SNN architecture of ML-AT. Although the ambulatory features are recorded daily in the dataset, every participant recorded a single value for each IGTB in the overall study (Section 3). So, we duplicated the IGTB features for each input sample of the model per participant in order to implement the model. To avoid biasing the models' performance and address any data leakage while detecting positive and negative affect, we have removed both trait positive affect and trait negative affect among the trait characteristics during the implementation of both of the detection models. The training process of ML-AT has followed the same steps and number of iterations as the one in ML-A and is also outlined in Algorithm 1.

4.2.2. Separate metric learning models with ambulatory and trait features for each IGTB cluster of participants (ML-C)

To further explore ways to personalize the model, we developed clusters based on IGTB values recorded by the participants. IGTB

Algorithm 1 Algorithm Describing the Metric Learning with Ambulatory Features (ML-A) model and theMetric Learning with Ambulatory and Trait features (ML-AT) model.

Input = **folds** ($\mathcal{F}$), **participants** ($\mathcal{N}$), **features** X **and labels** ($\mathcal{Y}$)
Output = **Predicted label** ($\bar{\mathcal{Y}}$)

{Creating folds from participants, one fold for testing, one fold for validation, remaining for training. Samples are taken from the training set. Hyper-parameter tuning and early stopping are done using the validation set. During each for loop for fold, testing is done.}
for $f \leftarrow 1$ *to* F **do**
  for $n \leftarrow 1$ *to* N **do**
    $\mathbf{x}_{nm}, \mathbf{x}_{nm'} \leftarrow$ *formed pairs within each participant* (*similar*/*dissimilar*) *with* m *and* m' *samples where* $m \neq m'$
    $y_{nm}, y_{nm'} \leftarrow$ *original labels for* m *and* m' *samples*
    $\hat{y}_{nm}, \hat{y}_{nm'} \leftarrow$ *normalized labels for* m *and* m' *samples*
  end for
  $\mathbf{W} \leftarrow$ *randomly initialize weights for* $\mathbf{f}_\mathbf{W}$ *layer*(*s*)
  $\mathbf{V} \leftarrow$ *randomly initialize weights for* $\mathbf{g}_\mathbf{V}$ *layer*
  while ***not*** *early stopping* **do**
    update $\mathbf{W}$ & $\mathbf{V}$ *based on* l_F & l_R *where*,
      $l_F = \sum_n \sum_{m \neq m'} |(\| \mathbf{f}_\mathbf{W}(\mathbf{x}_{nm}) - \mathbf{f}_\mathbf{W}(\mathbf{x}_{nm'})\|_2 - \|\hat{y}_{nm} - \hat{y}_{nm'}\|_2)|$
      $l_R = \sum_n \sum_m \| \mathbf{g}_\mathbf{V}(\mathbf{f}_\mathbf{W}(\mathbf{x}_{\mathbf{nm}})) - y_{nm}\|_2^2$
      where $\mathbf{x}_{nm}$*includes ambulatory features for ML-A both ambulatory and trait features for ML-AT*
  end while
end for

questionnaires capture participants' trait characteristics, therefore this type of clustering stratifies participants based on their psychological characteristics, which are likely to impact the bio-behavioral recordings conditioned on the considered outcomes (36, 77–79). Since the dimension of IGTB is high (14 IGTBs), we reduced the dimension by finding the first 3 principal components of the IGTB feature space through principal component analysis (PCA) (101). Following that, clustering was performed via the K-means algorithm (102), where we decided upon the value of K by optimizing cluster distortion and inertia via the elbow method per (103). Here, *distortion* is the average of the Euclidean squared distance from each sample to the centroid of its assigned cluster, and *inertia* is the sum of squared distances of samples to their closest cluster center. The K value corresponding to the elbow (i.e., "corner") of the graph is taken as the optimum K, where the distortion and inertia values do not change significantly for higher K values. Since the clustering algorithm does not show a clear elbow pattern in **Figure 4** for our dataset, we found the optimal K to be $K = 4$ for our data empirically after experimenting with different K values ranging from 2–4 based on the performance of the validation set. The clusters were formed using only the training data to avoid biasing the output of the model as well as avoiding any data leakage. The K-means model fitted on the training data was used to extract the clusters for validation and test data, and the ML-AT models (**Section 4.2.1**) were trained on each cluster separately. **Figure 5** depicts PCA1 and PCA2 of the 4 clusters in different colors formed using the training dataset developed

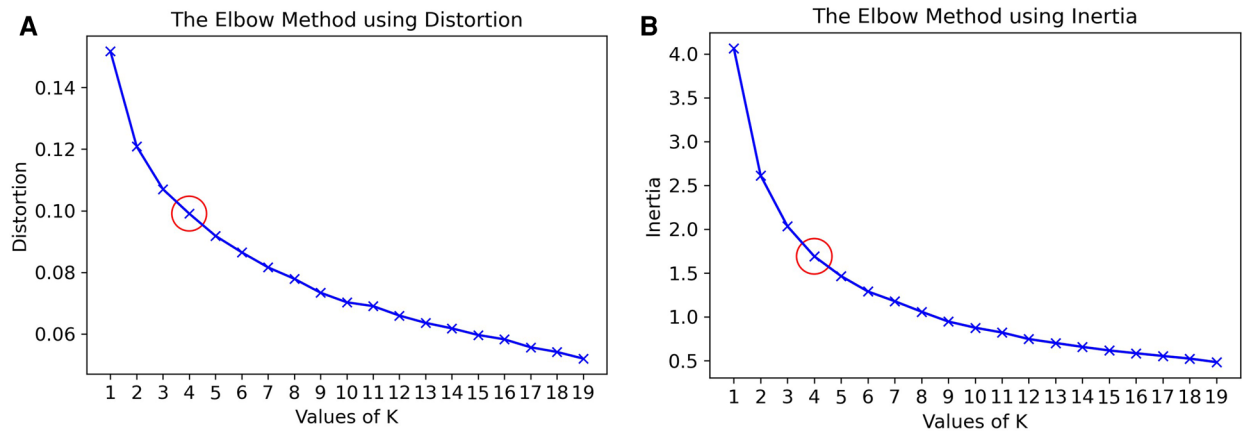


FIGURE 4
The Elbow method to select value of K in K-means algorithm to for clusters for the Siamese Model for Clustered IGTB. (A) Based on distortion and (B) based on inertia.

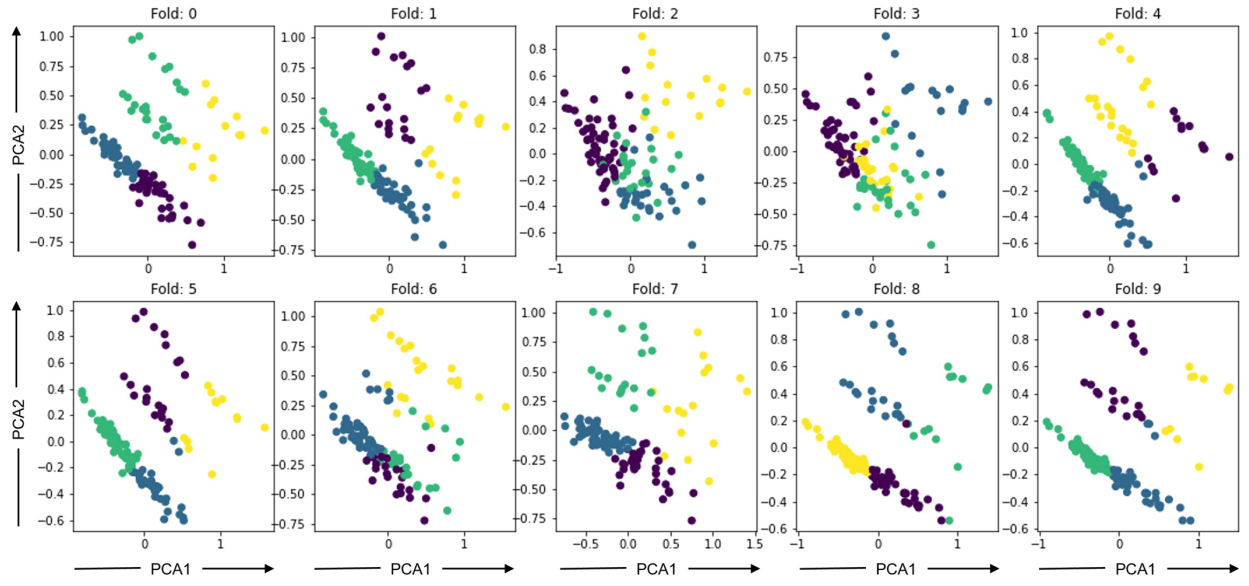


FIGURE 5
Clusters formed using the K-Means clustering method based on IGTB values where x axis is the principal component 1 (PCA1) and y axis is the principal component 2 (PCA2).

while iterating through the 10 cross-validation folds. We refer to the new cluster-based models as the separate metric learning models with ambulatory and trait features for each IGTB cluster of participants (ML-C), whose basic algorithm is described in **Algorithm 2**. The main difference between **Algorithms 1** and **2** is, one additional loop is introduced inside the outer loop looping through F. This additional loop loops through the clusters C, and inside the cluster, similar to **Algorithm 1**, two loops are formed to develop pairs (cluster based in this case) and training the model.

4.3. Experimental framework

The SNN, as shown in **Figure 3A**, has two parts: the shared f_W layers and the g_V layer. The number of layers in f_W is a hyperparameter with values 3–5, each layer containing 64 neurons, where g_V contains a single layer with a single node. We used ℓ_2 regularization (regularization value = 0.0001) on each layer and dropout between two layers for the purpose of proper regularization. The dropout value is another hyperparameter in our model with values ranging from 0.1–0.3. To avoid overfitting the model, we utilized early stopping based on the Pearson correlation coefficient of the validation data with patience as a hyperparameter (range: 3, 5, 10). The other hyperparameter we considered is the batch size (128, 256, 512). We optimized our models using stochastic gradient descent (SGD) (104) algorithm.

We used Bayesian optimization using Gaussian processes for hyperparameter tuning. The *gp_minimize* class of the *scikit-optimize* library was used for this purpose. The algorithm was run 200 times to ensure convergence while keeping the function to minimize over the gaussian prior at negative expected improvement. The whole process was repeated for all the folds developed for each of the constructs, keeping the random seed fixed to a constant (constant = 42) for a fair comparison across different runs.

After we fixed the participants and tuned the hyperparameters for each fold of the cross-validation loop on a fixed random seed, we saved this information and used it to build models, as well as to develop the train, validation, and test datasets for all the folds of the cross-validation loops separately. We ran these final models 20 times each with non-fixed random seeds ensuring that the model parameters were initialized differently during each run. As a result, we could explore different outcomes of the models for different initialization, thus understanding the overall performance potential of our models. Finally, we report the average of those outcomes and present the average correlation along with the standard deviation of the average value in **Figure 6**. It is important to note that, we removed the negative affect and positive affect IGTB features while training the ML-AT and ML-C models for the negative affect and positive affect constructs to avoid any data leakage (**Section 4.2.1**), reducing the number of IGTB features for these two constructs to 12 while producing the **Figures 6C,D**.

For both ML-A and ML-AT models, we have around 34 thousand pairs for training purposes. The number of similar and dissimilar samples is unbalanced among these pairs. To mitigate the effect of unbalanced pairing while training the models, we weighted the loss function based on the weight of similar and dissimilar pairs of a given cross-validation setting, which gave more weight to the set with fewer pairs. For the ML-C model, we have around $10k$ pairs for each cluster.

4.4. Baseline methods

We used two baseline methods to compare the performance of the SNN model-based ablation study described in **Section 4.3**. To

Algorithm 2 Algorithm Describing the Separate Metric Learning Models with Ambulatory and Trait Features for Each IGTB Cluster of Participants Model (ML-C).

Input = **folds** (F), **participants** (N), **clusters** (C), **features** ($\mathcal{X}$) **and labels** ($\mathcal{Y}$)
Output = **Predicted label** ($\bar{\mathcal{Y}}$)

{Creating folds from participants, one fold for testing, one fold for validation, remaining for training. Samples are taken from the training set. Hyper-parameter tuning and early stopping are done using the validation set. During each for loop for fold, testing is done.}
for $f \leftarrow 1$ *to* F **do**
  forming 4 *clusters based first* 3 *PCA derived from IGTB values of the training data*
  for $c \leftarrow 1$ *to* C **do**
    for $n \leftarrow 1$ *to* N **do**
      $\{\mathbf{x}_{nm}, \mathbf{x}_{nm'}\}_c \leftarrow$ *formed similar/dissimilar pairs from samples* m *and* $m'(m \neq m')$ *of participant* n *from cluster* c
      $\{y_{nm}, y_{nm'}\}_c \leftarrow$ *original labels of samples* m *and* m' *of participant* n *from cluster* c
      $\{\hat{y}_{nm}, \hat{y}_{nm'}\}_c \leftarrow$ *normalized labels of samples* m *and* m' *of participant* n *from cluster* c
    end for
    $\mathbf{W}_c \leftarrow$ *randomly initialize weights for* $\mathbf{f_W c}$*layer*(s) *of model corresponding to cluster* c
    $\mathbf{V}_c \leftarrow$ *randomly initialize weights for* $\mathbf{g_V c}$ *layer of model corresponding to cluster* c
    while not *early stopping* **do**
      update $\mathbf{W}_c$ $\mathbf{V}_c$ *based on* l_{F_c} l_{R_c} *where,*
        $l_{F_c} = \sum_n \sum_{m \neq m'} |(\| \mathbf{f_W c}(\mathbf{x}_{nm}) - \mathbf{f_W c}(\mathbf{x}_{nm'})\|_2 - \|\hat{y}_{nm} - \hat{y}_{nm'}\|_2)|$
        $l_{R_c} = \sum_n \sum_m \|g_{\mathbf{V}} \mathbf{c}(\mathbf{f_W c}(\mathbf{x}_{nm})) - y_{nm}\|_2^2$
        where $\mathbf{x}_{nm}$ *includes both ambulatory and trait features*
    end while
  end for
end for

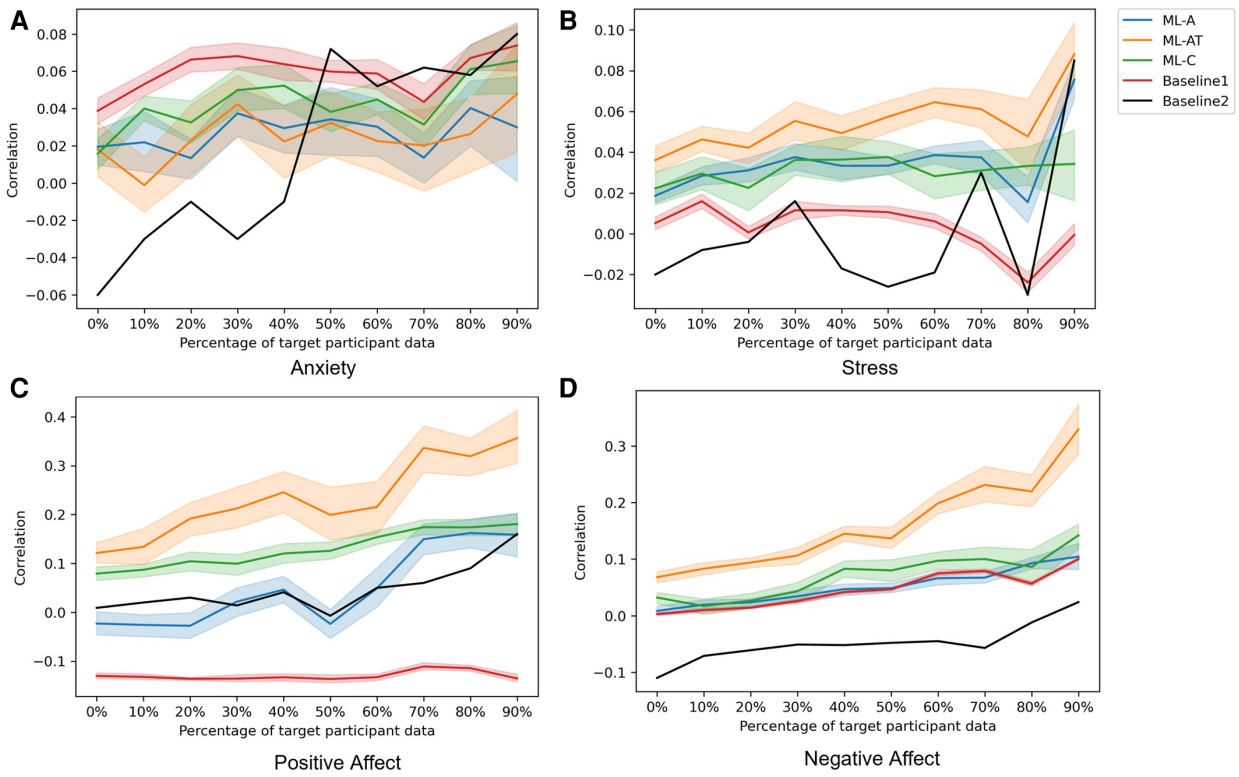


FIGURE 6
Average Pearson's correlation coefficient computed over 20 iterations of the considered models, each obtained with a different random restart. The *x*-axis corresponds to the different percentages of data from the target participants used for fine-tuning the models. The shaded shape around the solid line represents the standard deviation from the 20 iterations. Different colors representing different models as- Orange: ML-AT, Blue: ML-A, Green: ML-C, Red: Baseline1, Black: Baseline2.

train and evaluate the baselines, we followed the 10-fold stratified cross-validation method with the same 10 folds used in our proposed methods, following the same participants in each fold, the same hyperparameter for the neural network model construction and training, a similar number of trainable network parameters as the metric learning models (around 12k), and the models were fine-tuned with the same transfer learning technique as used for the proposed models. The ambulatory features were used as the input to the models of the baselines. The first baseline, named Baseline1, uses a feed-forward neural network (FNN), similar to the $\mathbf{g_V}$ layer of the SNN we propose, except Baseline 1 uses the number of layers equivalent to our proposed SNN models instead of the 1 layer $\mathbf{g_V}$. This baseline represents a non-personalized model where the ambulatory features serve as the input to the model, instead of the embedding learned through the metric learning process. Our objective in developing such a baseline is to examine whether the learned embedding outperforms the original feature space in predicting different constructs as well as how effective the personalized models are over the non-personalized models in this particular dataset.

As our second baseline, named Baseline2, we utilized the semi-supervised maximum independence domain adaptation (SMIDA) algorithm (105). The SMIDA algorithm proposes a way to reduce the discrepancy between data distribution of different domains. Domain discrepancy arrives in different forms, such as data collected using different devices, in different scenarios. In our case, all the participants' data were collected using the same types of devices and in a similar scenario. But, inter-individual variation in the data prompted us to design the SMIDA model by considering each participant as a different domain and reducing the domain discrepancy in the feature space through a projection matrix. The algorithm proposes an unsupervised and semi-supervised method of reducing domain discrepancy. We followed the semi-supervised method by keeping the test participants' data masked while learning the projection matrix. Among the kernels developed for creating the projection matrix, we opted for the linear one in our experiments. The projected matrix has been used as input to a linear regression model that estimated the MGT output.

5. Results

Although our highly skewed dataset is very suitable for classification experiments (**Figure 1**), instead of forming a classification problem from our dataset, we focused on doing a regression analysis to predict the original labels through our model. Because performing regression analysis on the constructs has given us the ability to avoid artificially discretizing the

predicted labels. Moreover, this improves the resolution of detection of the constructs thus making the models better suitable for designing JITAIs. We have used Pearson's correlation coefficient to evaluate our proposed models and baselines. In this section, we report the average and standard deviation of Pearson's correlation coefficient between the original and predicted MGTs across 20 iterations (i.e., different random initializations of the models) when using an increasing percentage of samples from the target participants for fine-tuning (**Figure 6**). Different colors in **Figure 6** represent the different models described in **Section 4**. According to our findings, the ML-AT model works the best in all the constructs we examined except for anxiety. This finding demonstrates that introducing trait information helps in improving the performance of the models and yields better accuracy when estimating daily constructs. Trait characteristics were not employed as sole predictors, because our goal is to estimate participants' states at the daily level. This requires behavioral information recorded via physiological and acoustic signals, allowing to capture of momentary fluctuations that cannot be recorded via the trait scores. Although our proposed models do not outperform baselines while predicting anxiety, the performance of the models is not significantly lower compared to the baseline methods for anxiety (**Table 3**). Apart from positive affect, all other constructs have severely skewed distribution (**Figure 1**), which might also affect the overall performance of the models. The second-best performing model for positive and negative affect is the ML-C (**Figures 6C,D**), while ML-C and ML-A perform similarly in the case of stress (**Figure 6B**). We have performed experiments keeping all 14 IGTB features for the positive and negative affect, which produces similar graphs as in **Figures 6C,D**. We further observe an increasing trend in performance for all the models when using more data from the target participants for fine-tuning. Such a trend is expected as models trained using a portion of the data from the target participants learn their unique distribution, thus performing better in the test samples. The varying performance of the clustering algorithm (i.e., some clusters are better separated in some folds compared to other folds in **Figure 5**) appears to affect the overall performance of the ML-C model. For example, the Pearson's correlation coefficient for the positive affect construct for fold 1, which has separable clusters, is 0.28 ($p < 0.001$), while the same value decreases to 0.04 ($p > 0.05$) for fold 3 with less separable clusters.

Next, we perform one-tailed paired-sample t-tests between the ML-AT, ML-A, and ML-AT, Baseline1, to examine the statistical significance between the proposed approaches and the baseline methods (**Table 3**). Results are reported for three different percentages of samples from target participants in the training data (i.e., 0%, 40%, and 90%). The ML-AT model's performance for stress, negative affect, and positive affect is significantly higher ($p < 0.05$) compared to the ML-A and Baseline1 models. Baseline1 performs the best for anxiety, although the performance among different models becomes similar as we increase the percentage of data taken from our test dataset for fine-tuning the model.

Through our proposed metric learning approach, we train the network by learning the relative distance between features with respect to the behavioral constructs of interest. In this way, the embeddings learned through this process should be able to reliably estimate the behavioral constructs (**Figure 6**, **Table 3**), while they should not retain person-specific information. To test this hypothesis, we have used the original features and the embeddings learned by the ML-A to classify the different participants of the dataset. We have utilized a logistic regression (LR) model for this person identification task and have considered embeddings learned from the different percentages of samples from the target participant used for fine-tuning. The average accuracy is calculated across 20 iterations that correspond to random initializations of the LR models (**Table 4**). The embedding features learned by the ML-A model appear to retain less participant-dependent information compared to the original features, as reflected in the lower participant classification accuracies of the first model (e.g., 41.92%) compared to the second (e.g., 83.21%). The fact that the embeddings learned by the ML-A models depict reduced variability across participants might also be a potential reason why the proposed metric learning approaches outperform conventional models, which might not be able to fully disentangle participant-dependent and construct-dependent information.

We explored the effect of the different IGTBs on constructing the PCA dimensions. For this purpose, we plotted a correlation circle plot showing the correlation between different IGTB values with the first two PCA dimensions (**Figure 7**). PCA 1 and PCA 2 explain 31.96% and 16.61% of the total variance of the data, respectively. The large inter-individual differences played a significant role in reducing the amount of variability explained by the first two principal components. Through the unit circle, it is easy to compare the effect of different IGTBs on developing the PCA dimensions. From the figure, we can find that Tobacco use (Yes/ No), which reflects whether the person uses tobacco or not, has the largest contribution to building the PCA dimensions, but contributes almost equally to both the first and second PCA

TABLE 3 Results from one-tailed paired-sample t-test evaluating significant differences between the predicted values from ML-AT, ML-A and Baseline 1 in the form of average Pearson's correlation coefficient over 20 iterations between true momentary ground truth (MGT) and predicted values of different constructs utilizing different percentage (%) of target data for fine tuning purpose.

MGT	Percentage (%)	ML-AT vs ML-A	ML-AT vs Baseline1
Anxiety	0	$t(19) = -0.15, p > 0.05$	$t(19) = -2.5, p > 0.05$
	40	$t(19) = -0.6, p > 0.05$	$t(19) = -3.79, p < 0.001$
	90	$t(19) = 0.86, p > 0.05$	$t(19) = -1.6, p > 0.05$
Stress	0	$t(19) = 4.33, p < 0.001$	$t(19) = 8.89, p < 0.001$
	40	$t(19) = 2.72, p < 0.001$	$t(19) = 8.36, p < 0.001$
	90	$t(19) = 1.32, p > 0.05$	$t(19) = 10.44, p < 0.001$
Positive effect	0	$t(19) = 8.6, p < 0.001$	$t(19) = 21.5, p < 0.001$
	40	$t(19) = 7.61, p < 0.001$	$t(19) = 16.74, p < 0.001$
	90	$t(19) = 5.39, p < 0.001$	$t(19) = 17.42, p < 0.001$
Negative effect	0	$t(19) = 8.95, p < 0.001$	$t(19) = 13.27, p < 0.001$
	40	$t(19) = 11.58, p < 0.001$	$t(19) = 15.38, p < 0.001$
	90	$t(19) = 8.54, p < 0.001$	$t(19) = 9.76, p < 0.001$

TABLE 4 Person identification accuracy in percentage (%) over 20 iterations using original features and embedding learnt from the Siamese Model (ML-A) for different percentage of fine tuning data from different momentary ground truths (MGT).

MGT	Features	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
Anxiety	Original	83.21	81.28	79.66	77.96	76.11	74.61	71.29	60.78	53.93	15.86
	Embedded	41.92	38.71	36.44	35.06	30.79	28.63	27.13	18.86	17.75	10.06
Stress	Original	83.15	80.96	79.60	77.18	76.99	73.93	67.68	58.96	49.59	17.20
	Embedded	53.31	50.83	48.28	47.46	44.06	40.59	37.97	28.60	25.09	11.22
Positive effect	Original	83.11	81.08	79.71	76.98	74.92	73.55	69.78	57.40	48.03	15.85
	Embedded	47.10	45.03	42.16	41.86	37.59	35.05	32.71	25.40	24.17	13.58
Negative effect	Original	83.11	81.34	79.25	77.85	76.08	73.49	68.65	60.38	47.08	15.76
	Embedded	51.79	48.64	48.12	46.09	42.98	38.31	37.64	30.22	26.22	14.29

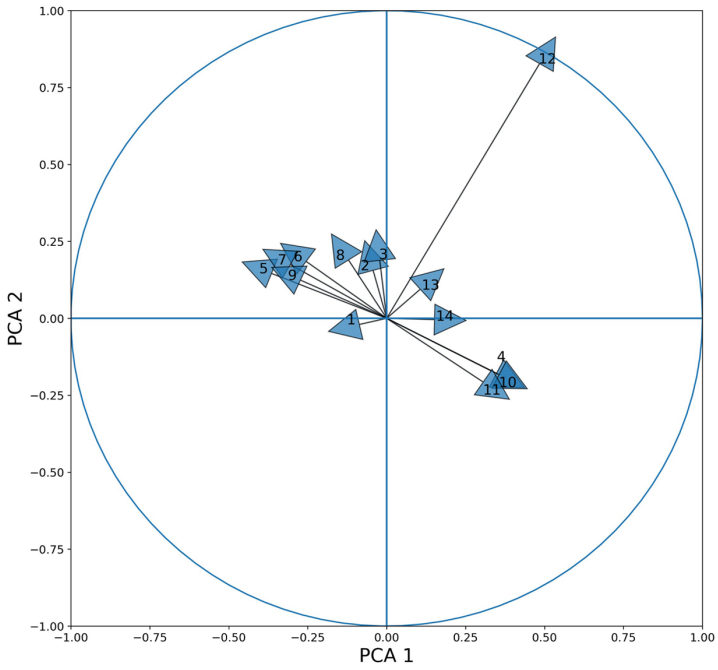
TABLE 5 Mean of normalized IGTB values across four clusters with ANOVA, Kruskal-Wallis, and t-test between clusters. Since most values are statistically significant, only non-significant values are shown using an asterisk.

Name of IGTB	Cluster				Statistics value							
	0	1	2	3	$F(3, 3596)^a$	$F(3, 3596)^b$	T-test between clusters					
							0 vs 1	0 vs 2	0 vs 3	1 vs 2	1 vs 3	2 vs 3
							$t(2018)$	$t(2234)$	$t(2647)$	$t(951)$	$t(1364)$	$t(1580)$
Organization citizenship behaviors	0.53	0.36	0.47	0.43	107.8	205.83	17.9	5.98	11.76	−8.87	−6.99	3.15
Fluid intelligence	0.58	0.62	0.59	0.63	9.56	49.61	−3.21	−2.37	−6.41	1.80*	−0.61*	−3.70
Crystallized intelligence	0.62	0.66	0.62	0.57	6.03	7.67	−3.19	0.30*	6.03	2.89	6.84	4.20
Emotional stability	0.24	0.66	0.30	0.48	1164.29	877.31	−50.84	−6.67	−44.30	36.30	22.69	−22.81
Conscientiousness	0.82	0.36	0.73	0.55	1363.63	973.16	53.78	12.51	35.38	−36.01	−18.21	18.88
Extraversion	0.65	0.35	0.63	0.41	422.14	621.88	31.84	2.34	36.34	−25.24	−6.25	24.97
Agreeableness	0.74	0.42	0.65	0.52	642.54	737.06	40.55	10.72	29.60	−24.14	−10.52	14.50
Openness	0.67	0.62	0.65	0.52	14.32	37.86	5.23	2.16	19.77	−3.12	8.13	13.96
Trait positive affect	0.70	0.41	0.59	0.50	416.55	604.60	30.35	13.83	27.14	−18.15	−8.64	11.82
Trait negative affect	0.15	0.59	0.16	0.33	1370.23	742.50	−32.42	−2.55	−25.84	29.87	17.98	−20.08
Trait anxiety	0.21	0.55	0.21	0.42	1078.97	798.20	−36.07	0.46*	−41.96	33.66	13.69	−33.52
Tobacco use (Yes/ No)	0	0.80	0.70	0.03	7399.80	2477.45	−61.42	−68.68	−7.25	6.11	57.43	62.59
Tobacco quantity	0	0.11	0.04	0	188.74	921.30	−9.08	−9.58	N/A	5.16	9.08	9.58
Sleep quality	0.22	0.38	0.25	0.30	148.56	179.32	−13.10	−4.24	−10.54	9.99	6.86	−5.09

[a]ANOVA
[b]Kruskal-Wallis
*$p > 0.05$ (not significant)

components. Fluid intelligence, crystallized intelligence, and openness contribute largely toward the second PCA dimension, while sleep quality contributes mostly to the first. Trait negative affect, trait anxiety, and emotional stability form a cluster contributing positively to the first PCA component and negatively to the second. On the other hand, extraversion, agreeableness, trait positive affect, organization citizenship behavior, form a different cluster contributing positively to the second PCA component and negatively to the first one.

In inspecting those clusters in more detail, **Table 5** further shows the mean of the normalized IGTB values computed for the participants of each cluster. Analysis of Variance (ANOVA), along with the Kruskal-Wallis test, indicates that all IGTB measures are significantly different across these four clusters (cluster 0 with 1651 samples, cluster 1 with 368 samples, cluster 2 with 584 samples, and cluster 3 with 997 samples). Upon close examination of this table, we can provide an additional description of the characteristics of the participants in each cluster. Since the Tobacco quantity value is zero in clusters 0 and 3, the t-test statistics in N/A in this case. Cluster 0 includes participants who have scored the highest on organization citizenship behaviors, conscientiousness, extraversion, agreeableness, openness, and positive affect. On the contrary, the cluster shows the lowest mean score for fluid intelligence, emotional stability, negative affect, trait anxiety, Tobacco use (Yes/ No) and use, and sleep quality. We note here that the lower the score for sleep quality, the better the overall sleep is. Thus, this cluster contains participants who do not use tobacco, have good sleep quality, have low anxiety, and are socially extroverted, highly agreeable, and open-minded. On the contrary, individuals assigned to cluster 1 appear to have opposite traits compared to the ones assigned to cluster 0 with the corresponding participants depicting the lowest values in organization citizenship behaviors, conscientiousness, extraversion, agreeableness, and positive affect; and the highest values in both fluid and crystallized intelligence, emotional stability, negative affect, trait anxiety, Tobacco use (Yes/No), and sleep quality. These suggest that cluster 1 includes participants with high intelligence, but lower overall well-being and higher negative traits. Clusters 2 and 3 appear to be "in the middle."



Number	Name of IGTB
1	Organization Citizenship Behaviors
2	Fluid intelligence
3	Crystalized intelligence
4	Negative emotionality
5	Conscientiousness
6	Extraversion
7	Agreeableness
8	Open mindedness
9	Trait anxiety
10	Tobacco status
11	Tobacco use (quantity)
12	Sleep quality

FIGURE 7
Correlation circle plot with a unit circle for PCA 1 and PCA 2 using different IGTB values.

Cluster 2 is more closely related to cluster 0. Individuals in cluster 2 have some of the lowest values for emotional stability and trait anxiety. These constructs are not significantly different between clusters 0 and 2 based on the post-hoc *t*-test analysis. Similarly, with the exception of the low values for crystallized intelligence and tobacco use, cluster 3 is more closely related to cluster 1. Individuals belonging to this cluster scored the lowest on openness and high on crystallized intelligence, emotional stability, and trait anxiety. Summarizing our findings, we can see that cluster 0 and cluster 2 are similar and include individuals with positive affectivity who tend to keep a relatively healthy lifestyle in terms of sleep and tobacco use. On the other hand, cluster 1 and cluster 3 are similar in that they include individuals with negative affectivity and relatively less healthy lifestyle.

6. Discussion

In this work, we have examined the effectiveness of metric learning algorithms in learning behavioral outcomes from ambulatory data, potentially addressing the inherent inter-individual variability that tends to hamper model performance. The performance of our proposed model has been evaluated on ambulatory data collected in an uncontrolled environment with respect to several constructs related to the mental and emotional well-being of healthcare workers, such as affect, stress, and anxiety. We have achieved personalization by utilizing the trait characteristics of the participants as an additional input to the models as well as clustering criteria for grouping the participants, followed by training separate metric learning models for each group. Personalization implemented by adding the IGTB features as an additional input to the model (i.e., ML-AT) outperforms the ML-A models that only include the ambulatory-based features suggesting the need for adding information about individuals' trait characteristics in the model. Moreover, our results have demonstrated that fine-tuning our model using a portion of data from the target participants can improve the model performance rather than training the model solely on samples from participants other than the target. We have also explored the role of different IGTBs on the resulting participant cluster. This provides us an additional intuition on the key participant traits to focus on in order to develop group-based models.

Studies among healthcare workers have shown that improvement of the overall workplace environment is possible through proper mindful interventions (106, 107). Healthcare workers are becoming more aware of their health, safety, and wellness, while issues associated with managing stress and promoting overall well-being (e.g., weight control) are becoming more and more pressing. Technology tools, such as intelligent ambulatory assessment via personalized models, could help in assisting healthcare workers for the same purpose. Prior work has found that healthcare workers view positively the integration of such smartphone-based technologies in their daily activities which could potentially lead to improving their well-being and mitigate the risk of depression (108), thus similar personalized machine learning models could be used as a foundation of such smartphone apps. The proposed metric learning model is quite lightweight (i.e., around only 12K trainable parameters) compared to the state-of-the-art convolution neural network (CNN) for mobile applications, MobileNetV3's 2.9 million parameters (109) and can utilize a small amount of collected data for training purposes, rendering it suitable for lightweight

app development of momentary psychological evaluation and intervention. This alternate way to assess the mental condition of healthcare workers has further the potential to yield higher adherence rates compared to conventionally-used self-reporting via ecological momentary assessment (EMA) since it only requires participants to wear their wearable devices and have their smartphone devices in close proximity.

Our proposed SNN and personalization techniques demonstrate improvements over methods that do not account for individual differences, however, the performance is yet insufficient for creating effective JITAIs. Utilizing a dataset collected in a real environment with multi-modal features has posed a unique challenge. Further, we have formulated the task as a regression problem, where the commonly used matrix related to tasks predicting similar constructs are accuracy and the F1 score (110–114). A recent review of studies classifying stress demonstrates that the overall accuracy of detecting stress through data collected from ambulatory devices (around 75%) is significantly lower than utilizing data collected in a laboratory setup (around 95%) (115, 116). From the distribution of the labels of different constructs (**Figure 1**) we observe that, apart from positive affect, all the other constructs have a severely skewed distribution. Especially, for anxiety, more than 50% data belongs to label 1, and for negative affect, almost 80% of the data belongs to label 5. This might be the underlying factor for the relatively better performance for predicting positive affect than the other constructs using our proposed models. In comparing our results with prior work, Pearson's correlation coefficients in estimating anxiety from ambulatory data were found to be similar to ours (cross-validated $R2$ of 0.06, permutation test $p < 0.001$) (117). Although Yan et al. (66) utilized the TILES-2018 dataset to predict the affective state of the participants via regression analysis, the method of personalization carried out in this work varies since the authors used data from participant's prior week's to predict the affect in future weeks. Moreover, the individual-specific models, trained using rescaled and transformed features based on individual-specific statistics (i.e. average and standard deviation) contrast significantly with our fold-wise cross-validation approach. Finally, along with the factors discussed above, the usage of contextual features and a relatively larger dataset (i.e., a combination of TILES-2018 and 2019, with 10-weeks data per participant as opposed to on average 3 weeks of data used in our study) could explain the overall better performance of this study. Overall, there is a shortage of prior work in ambulatory monitoring that uses similar evaluation metrics for our considered behavioral constructs. Another potential reason for the relatively low performance of our models lies in the fact that the self-reported labels used to train these models might not necessarily be collected in close temporal proximity to the occurrence of the daily event that led to the corresponding value at the self-report. For instance, a participant might have reported negative affect at the end of the day, but this report might have reflected an event that happened many hours before the self-report questionnaire was administered. The proximity between event and self-report could significantly impact the overall performance of the models. For example, using data collected in the wild from 606 individuals (67) achieved a 0.25 Spearman's correlation for stress detection. Apart from the fact that the dataset is relatively larger than the dataset we used in this study, this data also ensured that the participants would respond to the self-reports in close proximity to the event. This could potentially help in achieving relatively better performance. This lack of coordination between the event and the label could be potentially addressed via saliency detection methods, which will allow us to explore prominent temporal patterns in the data that might be indicative of changes in behavioral constructs.

Despite the encouraging results, our study presents the following limitation. So far, we have studied well-being measures that are aggregated over an entire day. As part of our future work, we will explore the dataset in finer time resolution than predicting our target constructs on a daily level, which will be the next step in developing an effective JITAI using our proposed approach. Our current MGT labels are recorded once daily limiting our ability to use the data in achieving this objective. Moreover, exploring patterns in data across a day instead of using the daily average value for predicting the constructs would help in pinpointing specific patterns which may trigger stress, anxiety, or affect a person both positively or negatively, thus improving upon our current findings. Another limitation of this study is the pre-defined similarity threshold between samples in our proposed metric learning approach. Determining the appropriate value of the similarity threshold, potentially for each participant separately, could have further improved our results and revealed additional insights regarding individual differences in our data. This study is also limited by a small dataset. To address this limitation, we plan to expand our current study in a cross-corpus manner extending to other publicly available datasets in this or similar domain, for example, TILES-2019 (65), Tesserae (118), and publicly available multimodal dataset on stress detection of nurses (119). Our results indicate that eliminating individual-specific differences from the data via metric learning can yield improved performance in detecting well-being outcomes. This can inspire the use of differential privacy approaches (e.g., using adversarial learning), which can potentially further eliminate individual-dependent attributes from our data and preserve, or even further attenuate, the behavioral attributes of interest.

7. Conclusion

Mental well-being is a crucial part of overall well-being and effective work performance. This paper examines a machine learning algorithm based on metric learning that contributes to assessing one's mental and emotional well-being unobtrusively through widely-available ambulatory sensing technology. We have proposed a lightweight model which utilizes a metric learning technique to learn the relative distance between similar and dissimilar samples and thereby reduce person-specific signal dependencies in the data. We have bolstered the personalization of our proposed approach by utilizing the participants' trait

characteristics as added features to the model and criteria for clustering the population of the dataset to form cluster-specific models. The performance of the proposed models has been compared against two baselines, one that has utilized a non-personalized learning technique via a model trained on all participants, and one that has integrated personalization via a domain adaptation technique rather than metric learning. Results from our experiments indicate that our proposed models outperform the baselines in most cases. Among the proposed approaches, models utilizing IGTB features perform better than those not using the IGTB features, highlighting the need for personalization. The proposed system paves the way for developing in-the-moment interventions that can promote a healthier workplace environment for healthcare workers and other professionals who work in similar high-paced high-stakes environments.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://tiles-data.isi.edu/homepage.

Ethics statement

The studies involving human participants were reviewed and approved by Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2017-17042800005. The patients/participants provided their written informed consent to participate in this study.

Author contributions

PP worked on developing models, preparing the data for models, result analysis, and writing. KM, AN, BMB, and SSN worked on data collection, supervision of model adoption, and manuscript edition. SSN and TC supervised the project along with writing the manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

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.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Landsbergis PA. Occupational stress among health care workers: a test of the job demands-control model. *J Organ Behav*. (1988) 9:217–39. doi: 10.1002/job.4030090303

2. Spoorthy MS. Mental health problems faced by healthcare workers due to the COVID-19 pandemic: a review. *Asian J Psychiatr*. (2020) 51:102119. doi: 10.1016/j.ajp.2020.102119

3. Saragih ID, Tonapa SI, Saragih IS, Advani S, Batubara SO, Suarilah I, et al. Global prevalence of mental health problems among healthcare workers during the COVID-19 pandemic: a systematic review, meta-analysis. *Int J Nurs Stud*. (2021) 121:104002. doi: 10.1016/j.ijnurstu.2021.104002

4. Rotenstein LS, Berwick DM, Cassel CK. Addressing well-being throughout the health care workforce: the next imperative. *JAMA*. (2022) 328:521–2. doi: 10.1001/jama.2022.12437

5. Chutiyami M, Cheong AM, Salihu D, Bello UM, Ndwiga D, Maharaj R, et al. COVID-19 pandemic, overall mental health of healthcare professionals globally: a meta-review of systematic reviews. *Front Psychiatry*. (2022) 12:2600. doi: 10.3389/fpsyt.2021.804525

6. Kawanishi C. Designing, operating a comprehensive mental health management system to support faculty at a university that contains a medical school, university hospital. *Seishin Shinkeigaku Zasshi*. (2016) 118:28–33.PMID: 27192789

7. Hamric AB, Epstein EG. A health system-wide moral distress consultation service: development, evaluation. In *HEC forum*. Vol. 29. Springer (2017). p. 127–43.

8. Wachter R, Goldsmith J. To combat physician burnout, improve care, fix the elecronic health record. *Harv Bus Rev*. (2018). Available at: https://hbr.org/2018/03/to-combat-physician-burnout-and-improve-care-fix-the-electronic-health-record.

9. Cunningham T, Çayir E. Nurse leaders employ contemplative practices to promote healthcare professional well-being, decrease anxiety. *J Nurs Adm*. (2021) 51:156–61. doi: 10.1097/NNA.0000000000000987

10. Bauer-Wu S, Fontaine D. Prioritizing clinician wellbeing: the university of Virginia's compassionate care initiative. *Glob Adv Health Med*. (2015) 4:16–22. doi: 10.7453/gahmj.2015.042

11. Weight CJ, Sellon JL, Lessard-Anderson CR, Shanafelt TD, Olsen KD, Laskowski ER. Physical activity, quality of life and burnout among physician trainees: the effect of a team-based, incentivized exercise program. In: *Mayo Clinic Proceedings*. Vol. 88. Elsevier (2013). p. 1435–42.

12. Zackoff M, Sastre E, Rodgers S. Vanderbilt wellness program: model and implementation guide. *MedEdPORTAL*. (2012) 8:9111. doi: 10.15766/mep_2374-8265.9111

13. Gordon JS. Mind-body skills groups for medical students: reducing stress, enhancing commitment,, promoting patient-centered care. *BMC Med Educ*. (2014) 14:1–8. doi: 10.1186/1472-6920-14-198

14. Klatt M, Steinberg B, Duchemin A-M. Mindfulness in motion (MIM): an onsite mindfulness based intervention (MBI) for chronically high stress work environments to increase resiliency, work engagement. *J Vis Exp*. (2015) 101:e52359. doi: 10.3791/52359

15. Perski O, Hébert ET, Naughton F, Hekler EB, Brown J, Businelle MS. Technology-mediated just-in-time adaptive interventions (JITAIS) to reduce harmful substance use: a systematic review. *Addiction*. (2022) 117:1220–41. doi: 10.1111/add.15687

16. Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, Tewari A, et al. Microrandomized trials: an experimental design for developing just-in-time adaptive interventions. *Health Psychol*. (2015) 34:1220. doi: 10.1037/hea0000305

17. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components, design principles for ongoing health behavior support. *Ann Behav Med*. (2018) 52:446–62. doi: 10.1007/s12160-016-9830-8

18. D'Mello SK. Automated mental state detection for mental health care. In *Artificial intelligence in behavioral, mental health care*. Elsevier (2016). p. 117–36.

19. Kaare KK, Otto T. Smart health care monitoring technologies to improve employee performance in manufacturing. *Procedia Eng*. (2015) 100:826–33. doi: 10.1016/j.proeng.2015.01.437

20. Shamian J, El-Jardali F. Healthy workplaces for health workers in Canada: knowledge transfer and uptake in policy and practice. *Healthc Pap*. (2007) 7:6. doi: 10.12927/hcpap.2007.18668

21. Kozey-Keadle S, Libertine A, Lyden K, Staudenmayer J, Freedson PS. Validation of wearable monitors for assessing sedentary behavior. *Med Sci Sports Exerc*. (2011) 43:1561–7. doi: 10.1249/MSS.0b013e31820ce174

22. Donaldson SI, Grant-Vallone EJ. Understanding self-report bias in organizational behavior research. *J Bus Psychol*. (2002) 17:245–60. doi: 10.1023/A:1019637632584

23. L'Hommedieu M, L'Hommedieu J, Begay C, Schenone A, Dimitropoulou L, Margolin G, et al. Lessons learned: recommendations for implementing a longitudinal study using wearable and environmental sensors in a health care organization. *JMIR Mhealth Uhealth*. (2019) 7:e13305. doi: 10.2196/13305

24. Jones M, Johnston D. Understanding phenomena in the real world: the case for real time data collection in health services research. *J Health Serv Res Policy*. (2011) 16:172–6. doi: 10.1258/jhsrp.2010.010016

25. Jesus G, Casimiro A, Oliveira A. A survey on data quality for dependable monitoring in wireless sensor networks. *Sensors*. (2017) 17:2010. doi: 10.3390/s17092010

26. Yan K, Kou L, Zhang D. Learning domain-invariant subspace using domain features and independence maximization. *IEEE Trans Cybern*. (2017) 48:288–99. doi: 10.1109/TCYB.2016.2633306

27. Castellana A, Carullo A, Astolfi A, Puglisi GE, Fugiglando U. Intra-speaker and inter-speaker variability in speech sound pressure level across repeated readings. *J Acoust Soc Am*. (2017) 141:2353–63. doi: 10.1121/1.4979115

28. Quer G, Gouda P, Galarnyk M, Topol EJ, Steinhubl SR. 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*. (2020) 15:e0227709. doi: 10.1371/journal.pone.0227709

29. Chaspari T, Timmons AC, Margolin G. Population-specific and personalized (PSP) models of human behavior for leveraging smart and connected data. *Smart data*. Chapman and Hall/CRC (2019). P. 243–58.

30. Fiorini L, Cavallo F, Dario P, Eavis A, Caleb-Solly P. Unsupervised machine learning for developing personalised behaviour models using activity data. *Sensors*. (2017) 17:1034. doi: 10.3390/s17051034

31. Khademi A, El-Manzalawy Y, Buxton OM, Honavar V. Toward personalized sleep-wake prediction from actigraphy. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE (2018). p. 414–7.

32. Paromita P, Chaspari T, Sajjadi S, Das A, Mortazavi BJ, Gutierrez-Osuna R. Personalized meal classification using continuous glucose monitors. In: Glowacka D, Krishnamurthy VR, editors. *Joint Proceedings of the (ACM) (IUI) 2021 Workshops co-located with 26th (ACM) Conference on Intelligent User Interfaces (ACM) (IUI) 2021)*, College Station, United States, April 13–17, 2021. Vol. 2903. CEUR-WS.org (2021). https://ceur-ws.org/Vol-2903/IUI21WS-HEALTHI-10.pdf.

33. Koldijk S, Neerincx MA, Kraaij W. Detecting work stress in offices by combining unobtrusive sensors. *IEEE Trans Affect Comput*. (2018) 9:227–39. doi: 10.1109/TAFFC.2016.2610975

34. Taylor SA, Jaques N, Nosakhare E, Sano A, Picard R. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Trans Affect Comput*. (2020) 11(2):200–13. doi: 10.1109/TAFFC.2017.2784832

35. Gujral A, Chaspari T, Timmons AC, Kim Y, Barrett S, Margolin G. Population-specific detection of couples' interpersonal conflict using multi-task learning. In: *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. New York, NY, United States: Association for Computing Machinery (2018). p. 229–33.

36. Gupta K, Gujral A, Chaspari T, Timmons AC, Han S, Kim Y, et al. Sub-population specific models of couples' conflict. *ACM Trans Int Technol*. (2020) 20:1–20. doi: 10.1145/3372045

37. Zen G, Porzi L, Sangineto E, Ricci E, Sebe N. Learning personalized models for facial expression analysis and gesture recognition. *IEEE Trans Multimed*. (2016) 18:775–88. doi: 10.1109/TMM.2016.2523421

38. Fan H, Zheng L, Yan C, Yang Y. Unsupervised person re-identification: clustering and fine-tuning. *ACM Trans Multimed Comput Commun Appl*. (2018) 14:1–18. doi: 10.1145/3243316

39. LiKamWa R, Liu Y, Lane ND, Zhong L. Moodscope: building a mood sensor from smartphone usage patterns. In: *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*. New York, NY, United States: Association for Computing Machinery (2013). p. 389–402.

40. Hilliard N, Phillips L, Howland S, Yankov A, Corley CD, Hodas NO. Few-shot learning with metric-agnostic conditional embeddings [Preprint] (2018). Available at: http://arxiv.org/1802.04376.

41. Motiian S, Piccirilli M, Adjeroh DA, Doretto G. Unified deep supervised domain adaptation and generalization. In: *Proceedings of the IEEE International Conference on Computer Vision*. IEEE (2017). p. 5715–25.

42. Mundnich K, Booth BM, Feng T, Girault B, L'hommedieu J, Justin W, et al. TILES-2018, a longitudinal physiologic and behavioral data set of hospital workers. *Sci Data*. (2020) 7(1):1–26. doi: 10.1038/s41597-020-00655-3

43. Jacobson NC, Summers B, Wilhelm S. Digital biomarkers of social anxiety severity: digital phenotyping using passive smartphone sensors. *J Med Internet Res*. (2020) 22:e16875. doi: 10.2196/16875

44. Fahrenberg J, Myrtek M, Pawlik K, Perrez M. Ambulatory assessment-monitoring behavior in daily life settings. *Eur J Psychol Assess*. (2007) 23:206–13. doi: 10.1027/1015-5759.23.4.206

45. Sano A, Picard RW. Stress recognition using wearable sensors, mobile phones. In: *2013 Humaine Association Conference on Affective Computing, Intelligent Interaction*. IEEE (2013). p. 671–6.

46. Hunasgi S, Koneru A, Rudraraju A, Manvikar V, Vanishree M. Stress recognition in dental students using smartphone sensor, a software: a pilot study. *J Oral Maxillofac Pathol*. (2018) 22:314. doi: 10.4103/jomfp.JOMFP_168_18

47. Bogomolov A, Lepri B, Ferron M, Pianesi F, Pentland A. Daily stress recognition from mobile phone data, weather conditions, individual traits. In: *Proceedings of the 22nd ACM international conference on Multimedia*. New York, NY, United States: Association for Computing Machinery (2014). p. 477–86.

48. Epstein DH, Tyburski M, Kowalczyk WJ, Burgess-Hull AJ, Phillips KA, Curtis BL, et al. Prediction of stress and drug craving ninety minutes in the future with passively collected GPS data. *NPJ Digit Med*. (2020) 3:1–12. doi: 10.1038/s41746-020-0234-6

49. Kolenik T. Methods in digital mental health: smartphone-based assessment and intervention for stress, anxiety, and depression. In: *Integrating Artificial Intelligence and IoT for Advanced Health Informatics: AI in the Healthcare Sector*. Springer (2022). p. 105–28.

50. Fukazawa Y, Ito T, Okimura T, Yamashita Y, Maeda T, Ota J. Predicting anxiety state using smartphone-based passive sensing. *J Biomed Inform*. (2019) 93:103151. doi: 10.1016/j.jbi.2019.103151

51. Cai L, Boukhechba M, Wu C, Chow PI, Teachman BA, Barnes LE, et al. State affect recognition using smartphone sensing data. In: *Proceedings of the 2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies*. New York, NY, United States: Association for Computing Machinery (2018). p. 120–5.

52. Ahn J, Han R. Personalized behavior pattern recognition and unusual event detection for mobile users. *Mobile Inf Syst*. (2013) 9:99–122. doi: 10.1155/2013/360243

53. Bradley AH, Howard AL. Stress and mood associations with smartphone use in university students: a 12-week longitudinal study. *Clin Psychol Sci*. (2023) 0:21677026221116889. doi: 10.1177/21677026221116889

54. Diaz KM, Krupka DJ, Chang MJ, Peacock J, Ma Y, Goldsmith J, et al. Fitbit®: an accurate, reliable device for wireless physical activity tracking. *Int J Cardiol*. (2015) 185:138. doi: 10.1016/j.ijcard.2015.03.038

55. Yu H, Itoh A, Sakamoto R, Shimaoka M, Sano A. Forecasting health, wellbeing for shift workers using job-role based deep neural network. In: *Wireless Mobile Communication, Healthcare: 9th EAI International Conference, MobiHealth 2020, Virtual Event*; 2020 Nov 19; Proceedings. Springer (2021). p. 89–103.

56. Lawanont W, Mongkolnam P, Nukoolkit C, Inoue M. Daily stress recognition system using activity tracker, smartphone based on physical activity and heart rate data. In: *International Conference on Intelligent Decision Technologies*. Springer (2018). p. 11–21.

57. Yu H, Sano A. Passive sensor data based future mood, health and stress prediction: user adaptation using deep learning. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE (2020). p. 5884–7.

58. LiKamWa R, Liu Y, Lane ND, Zhong L. Can your smartphone infer your mood. In *PhoneSense Workshop*. PhoneSense workshop (2011). p. 1–5.

59. Bachmann A, Klebsattel C, Budde M, Riedel T, Beigl M, Reichert M, et al. How to use smartphones for less obtrusive ambulatory mood assessment and mood recognition. In: *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. PhoneSense workshop (2015). p. 693–702.

60. Schmid RF, Thomas J. The interactive effects of heart rate variability and mindfulness on indicators of well-being in healthcare professionals' daily working life. *Int J Psychophysiol* (2021) 164:130–8. doi: 10.1016/j.ijpsycho.2021.01.012

61. Sano A, Taylor S, McHill AW, Phillips AJ, Barger LK, Klerman E, et al. Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: observational study. *J Med Internet Res*. (2018) 20:e9410. doi: 10.2196/jmir.9410

62. Egilmez B, Poyraz E, Zhou W, Memik G, Dinda P, Alshurafa N. Ustress: understanding college student subjective stress using wrist-based passive sensing. In: *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE (2017). p. 673–8.

63. Wang R, Wang W, DaSilva A, Huckins JF, Kelley WM, Heatherton TF, et al. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proc ACM Interact Mob Wearable Ubiquitous Technol*. (2018) 2:1–26. doi: 10.1145/3191775

64. Gaballah A, Tiwari A, Narayanan S, Falk TH. 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 (2021). p. 8348–52.

65. Yau JC, Girault B, Feng T, Mundnich K, Nadarajan A, Booth BM, et al. Tiles-2019: a longitudinal physiologic and behavioral data set of medical residents in an intensive care unit. *Sci Data*. (2022) 9:536. doi: 10.1038/s41597-022-01636-4

66. Yan S, Hosseinmardi H, Kao H-T, Narayanan S, Lerman K, Ferrara E. Estimating individualized daily self-reported affect with wearable sensors. In: *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE (2019). p. 1–9.

67. Booth BM, Vrzakova H, Mattingly SM, Martinez GJ, Faust L, D'Mello SK. Toward robust stress prediction in the age of wearables: modeling perceived stress in a longitudinal study with information workers. *IEEE Trans Affect Comput*. (2022) 13:2201–17. doi: 10.1109/TAFFC.2022.3188006

68. Gonzalez-Carabarin L, Castellanos-Alvarado EA, Castro-Garcia P, Garcia-Ramirez MA. Machine Learning for personalised stress detection: Inter-individual variability of EEG-ECG markers for acute-stress response. *Comput Methods Programs Biomed*. (2021) 209:106314. doi: 10.1016/j.cmpb.2021.106314

69. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized nutrition by prediction of glycemic responses. *Cell*. (2015) 163:1079–94. doi: 10.1016/j.cell.2015.11.001

70. Zhong J, Liu L, Wei Y, Luo D, Sun L, Lu Y. Personalized activity recognition using molecular complex detection clustering. In: *2014 IEEE 11th International Conference on Ubiquitous Intelligence and Computing and 2014 IEEE 11th International Conference on Autonomic and Trusted Computing and 2014 IEEE 14th International Conference on Scalable Computing and Communications and Its Associated Workshops*. IEEE (2014). p. 850–4.

71. Amjad F, Khan MH, Nisar MA, Farid MS, Grzegorzek M. A comparative study of feature selection approaches for human activity recognition using multimodal sensory data. *Sensors* (2021) 21(7):2368. doi: 10.3390/s21072368

72. Sun X, Kashima H, Ueda N. Large-scale personalized human activity recognition using online multitask learning. *IEEE Transactions on Knowledge and Data Engineering* (2012) 25(11):2551–63. doi: 10.1109/TKDE.2012.246

73. De Santos A, Sánchez-Avila C, Guerra-Casanova J, Bailador-Del Pozo G. Real-time stress detection by means of physiological signals. In: *Recent application in biometrics*. InTech (2011).

74. Kallus N. Recursive partitioning for personalization using observational data. In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. JMLR.org (2017). p. 1789–98.

75. Bertsimas D, Kallus N, Weinstein AM, Zhuo YD. Personalized diabetes management using electronic medical records. *Diabetes Care*. (2017) 40:210–7. doi: 10.2337/dc16-0826

76. Zenonos A, Khan A, Kalogridis G, Vatsikas S, Lewis T, Sooriyabandara M. Healthyoffice: mood recognition at work using smartphones and wearable sensors. In: *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. IEEE (2016). p. 1–6.

77. Kuo T-H, Ho L-A. Individual difference and job performance: the relationships among personal factors, job characteristics, flow experience, and service quality. *J Soc Behav Pers*. (2010) 38:531–52. doi: 10.2224/sbp.2010.38.4.531

78. Ramsay DS, Woods SC. Clarifying the roles of homeostasis and allostasis in physiological regulation. *Psychol Rev*. (2014) 121:225. doi: 10.1037/a0035942

79. Schneiderman N, Ironson G, Siegel SD. Stress and health: psychological, behavioral, and biological determinants. *Annu Rev Clin Psychol*. (2005) 1:607–28. doi: 10.1146/annurev.clinpsy.1.102803.144141

80. Yadav M, Sakib MN, Feng K, Chaspari T, Behzadan A. Virtual reality interfaces and population-specific models to mitigate public speaking anxiety. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE (2019). p. 1–7.

81. Nguyen H, Ruiz C, Wilson V, Strong D, Djamasbi S. Using personality traits and chronotype to support personalization and feedback in a sleep health behavior change support system. In: *Proceedings of the 51st Hawaii International Conference on System Sciences*. AIS Electronic Library (AISeL) (2018).

82. Tondello GF, Arrambide K, Ribeiro G, Cen AJ-l., Nacke LE. "I don't fit into a single type": a trait model and scale of game playing preferences. In: *IFIP Conference on Human-Computer Interaction*. Springer (2019). p. 375–95.

83. Halperin E, Schori-Eyal N. Towards a new framework of personalized psychological interventions to improve intergroup relations and promote peace. *Soc Personal Psychol Compass*. (2020) 14:255–70. doi: 10.1111/spc3.12527

84. Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition. In: *ICML deep learning workshop*. Vol. 2; Lille Cambridge University Press. (2015).

85. Pflugfelder R. An in-depth analysis of visual tracking with Siamese neural networks [Preprint] (2017). http://arxiv.org/1707.00569.

86. Li MD, Chang K, Bearce B, Chang CY, Huang AJ, Campbell JP, et al. Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. *NPJ Digit Med*. (2020) 3:1–9. doi: 10.1038/s41746-020-0255-1

87. Arora P, Chaspari T. Exploring Siamese neural network architectures for preserving speaker identity in speech emotion classification. In: *Proceedings of the 4th International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*. New York, NY, United States: Association for Computing Machinery (2018). p. 15–8.

88. Lian Z, Li Y, Tao J, Huang J. Speech emotion recognition via contrastive loss under Siamese networks. In: *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and First Multi-Modal Affective Computing of Large-Scale Multimedia Data*. New York, NY, United States: Association for Computing Machinery (2018). p. 21–6.

89. Zhang C, Liu W, Ma H, Fu H. Siamese neural network based gait recognition for human identification. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE (2016). p. 2832–6.

90. Doumanoglou A, Balntas V, Kouskouridas R, Kim T-K. Siamese regression networks with efficient mid-level feature extraction for 3D object pose estimation [Preprint] (2016). Available at: http://arxiv.org/1607.02257.

91. Feng T, Nadarajan A, Vaz C, Booth B, Narayanan S. Tiles audio recorder: an unobtrusive wearable solution to track audio activity. In: *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*. New York, NY, United States: Association for Computing Machinery (2018). p. 33–8.

92. Fox S, Spector PE, Goh A, Bruursema K, Kessler SR. The deviant citizen: measuring potential positive relations between counterproductive work behaviour and organizational citizenship behaviour. *J Occup Organ Psychol*. (2012) 85:199–220. doi: 10.1111/j.2044-8325.2011.02032.x

93. Shipley WC, Gruber CP, Martin TA, Klein AM, *Shipley-2*. Los Angeles, CA: Western Psychological Services (2009).

94. Soto CJ, John OP. The next big five inventory (BFI-2): developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *J Pers Soc Psychol*. (2017) 113:117. doi: 10.1037/pspp0000096

95. Watson D, Clark LA. The PANAS-X: manual for the positive and negative affect schedule-expanded form. University of Iowa (1994).

96. Spielberger CD, Gorsuch RL, *State-trait anxiety inventory for adults: sampler set: manual, Tekst booklet and scoring key*. Sunnyvale: Consulting Psychologists Press (1983). https://shop.themyersbriggs.com/support/support.aspx

97. King BA, Mirza SA, Babb SD, GATS Collaborating Group and Others. A cross-country comparison of secondhand smoke exposure among adults: findings from the global adult tobacco survey (GATS). *Tob Control*. (2013) 22(4):e5–. doi: 10.1136/tobaccocontrol-2012-050582

98. Buysse DJ, Reynolds III CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. *Psychiatry Res*. (1989) 28:193–213. doi: 10.1016/0165-1781(89)90047-4

99. Mackinnon A, Jorm AF, Christensen H, Korten AE, Jacomb PA, Rodgers B. A short form of the positive and negative affect schedule: evaluation of factorial validity and invariance across demographic variables in a

community sample. *Pers Individ Dif*. (1999) 27:405–16. doi: 10.1016/S0191-8869(98)00251-7

100. Booth BM, Mundnich K, Feng T, Nadarajan A, Falk TH, Villatte JL, et al. Multimodal human and environmental sensing for longitudinal behavioral studies in naturalistic settings: framework for sensor selection, deployment, and management. *J Med Internet Res*. (2019) 21:e12832. doi: 10.2196/12832

101. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemometr Intell Lab Syst*. (1987) 2:37–52. doi: 10.1016/0169-7439(87)80084-9

102. Krishna K, Murty MN. Genetic K-means algorithm. *IEEE Trans Syst Man Cybern B (Cybern)*. (1999) 29:433–9. doi: 10.1109/3477.764879

103. Bholowalia P, Kumar A. EBK-means: a clustering technique based on elbow method and K-means in WSN. *Int J Comput Appl*. (2014) 105:17–24. doi: 10.5120/18405-9674

104. Amari S-i. Backpropagation and stochastic gradient descent method. *Neurocomputing*. (1993) 5:185–96. doi: 10.1016/0925-2312(93)90006-O

105. Yan K, Kou L, Zhang D. Domain adaptation via maximum independence of domain features [Preprint] (2016). Available at: http://arxiv.org/1603.04535.

106. Park J-H, Jung S-E, Ha D-J, Lee B, Kim M-S, Sim K-L, et al. The effectiveness of e-healthcare interventions for mental health of nurses: a prisma-compliant systematic review of randomized controlled trials. *Medicine*. (2022) 101:e29125–. doi: 10.1097/MD.0000000000029125

107. Jarden RJ, Sandham M, Siegert RJ, Koziol-McLain J. Strengthening workplace well-being: perceptions of intensive care nurses. *Nurs Crit Care*. (2019) 24:15–23. doi: 10.1111/nicc.12386

108. Kerst A, Zielasek J, Gaebel W. Smartphone applications for depression: a systematic literature review and a survey of health care professionals' attitudes towards their use in clinical practice. *Eur Arch Psychiatry Clin Neurosci*. (2020) 270:139–52. doi: 10.1007/s00406-018-0974-3

109. Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, et al. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE (2019). p. 1314–24.

110. Baumgartl H, Fezer E, Buettner R. Two-level classification of chronic stress using machine learning on resting-state EEG recordings. In *25th Americas Conference on Information Systems*. Association for Information Systems (2020).

111. Subhani AR, Mumtaz W, Saad MNBM, Kamel N, Malik AS. Machine learning framework for the detection of mental stress at multiple levels. *IEEE Access*. (2017) 5:13545–56. doi: 10.1109/ACCESS.2017.2723622

112. McGinnis RS, McGinnis EW, Hruschak J, Lopez-Duran NL, Fitzgerald K, Rosenblum KL, et al. Rapid anxiety and depression diagnosis in young children enabled by wearable sensors and machine learning. In: *2018 40th Annual International Conference of the IEEE Engineerin in Medicine and Biology Society (EMBC)*. IEEE (2018). p. 3983–6.

113. Pintelas EG, Kotsilieris T, Livieris IE, Pintelas P. A review of machine learning prediction methods for anxiety disorders. In: *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*. New York, NY, United States: Association for Computing Machinery (2018). p. 8–15.

114. Miranda-Correa JA, Patras I. A multi-task cascaded network for prediction of affect, personality, mood and social context using eeg signals. In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE (2018). p. 373–80.

115. Can YS, Chalabianloo N, Ekiz D, Fernandez-Alvarez J, Riva G, Ersoy C. Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches. *IEEE Access*. (2020) 8:38146–63. doi: 10.1109/ACCESS.2020.2975351

116. Booth BM, Vrzakova H, Mattingly SM, Martinez GJ, Faust L, D'Mello SK. Toward robust stress prediction in the age of wearables: modeling perceived stress in a longitudinal study with information workers. *IEEE Trans Affect Comput*. (2022) 13:2201–17. doi: 10.1109/TAFFC.2022.3188006

117. Boeke EA, Holmes AJ, Phelps EA. Toward robust anxiety biomarkers: a machine learning approach in a large-scale sample. *Biol Psychiatry*. (2020) 5:799–807. doi: 10.1016/j.bpsc.2019.05.018

118. Mattingly SM, Gregg JM, Audia P, Bayraktaroglu AE, Campbell AT, Chawla NV, et al. The tesserae project: large-scale, longitudinal, in situ, multimodal sensing of information workers. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, United States: Association for Computing Machinery (2019). p. 1–8.

119. Hosseini S, Gottumukkala R, Katragadda S, Bhupatiraju RT, Ashkar Z, Borst CW, et al. A multimodal sensor dataset for continuous stress detection of nurses in a hospital. *Sci Data*. (2022) 9:255. doi: 10.1038/s41597-022-01361-y