

Measuring affective state: Subject-dependent and -independent prediction based on longitudinal multimodal sensing

Lea Berkemeier, Wim Kamphuis, Anne-Marie Brouwer, Herman de Vries, Maarten Schadd, Jan Ubbo van Baardewijk, Hilbrand Oldenhuis, Rudolf Verdaasdonk, and Lisette van Gemert-Pijnen

Abstract—Current sensors offering passive and continuous monitoring of behavioral patterns potentially enable real-time affective state monitoring. Previous research on affective state prediction with multimodal sensing in daily life has shown only small-to-moderate effects. One reason for this limited success might be the large variability across individuals. Current research is often of short duration, preventing proper within-individual modeling. With an extensive longitudinal data collection of nine months, this research focuses on individual-level predictions of valence and arousal in daily life. Sixteen PhD candidates from the Netherlands provided data about their affective states (self-reported valence and arousal), physiology (Oura rings) and behavioral patterns (AWARE framework for mobile phone data). Supporting our hypothesis, subject-dependent random forest (RF) models significantly outperformed subject-independent leave-one-subject-out (LOSO) models in predicting self-reported valence and arousal. The subject-dependent models achieved an average Spearman's rho correlation of 0.28 [0.14-0.60] for valence and 0.36 [0.16-0.69] for arousal. In many cases, participants' a priori indicated informative sources matched with the feature importance. Making use of participants' self-knowledge might thus help to reduce the amount of data to be collected. For future work, longer-term changes in affective state and combinations of features for estimating real behavioral patterns should be explored.

Index Terms— Depression, Digital biomarkers, Individual, In-situ Studies, Longitudinal, Machine learning, Monitoring, Modeling and prediction, Physiological measures, Stress, Wearable sensors

Submitted on 26th of October 2023 to IEEE Transactions on Affective Computing for review. This work was supported by The Netherlands Organization for Applied Scientific Research (TNO). Corresponding author: L. Berkemeier.

L. Berkemeier is with the Research Group Digital Transformation, Hanze University of Applied Sciences, Zernikeplein 11, 9747 AS Groningen, The Netherlands, and with The Netherlands Organization for Applied Scientific Research (TNO), Kampweg 55, 3769 DE Soesterberg, The Netherlands and also with the Department of Psychology, Health and Technology, Centre for eHealth Research and Wellbeing, University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands (e-mail: lea.berkemeier@tno.nl).

W. Kamphuis, H. de Vries, and J.U. van Baardewijk are with The Netherlands Organization for Applied Scientific Research (TNO), Kampweg 55, 3769 DE Soesterberg, The Netherlands (e-mail: wim.kamphuis@tno.nl, herman.devries@tno.nl, jan_ubbo.vanbaardewijk@tno.nl).

I. INTRODUCTION

REAL-TIME monitoring of behavioral patterns becomes increasingly possible with the availability of sensor technologies and consumer-grade wearables. Behavioral patterns such as movement, physical activity or sleep can be quantified through continuous measurements that only require minimal user effort or input [1] [2]. Through this, small changes in a person's life or behavioral patterns can be monitored passively and detected automatically, which in turn may be used for the prevention of prospective problems and intervention of existing problems [3]. For example, [4] showed that mobile sensing data can be used to detect behavioral changes in patients with schizophrenia and found features predictive for oncoming psychotic relapses. Prospective problems, such as deteriorating mental health, can also play an essential role in the work-related context. For example, it can put employees at higher risk of developing mental disorders [5], burnout [6] or cardiovascular diseases [7]. Employees of different occupations can be considered at-risk, for instance, police officers or military service members due to repeated exposure to extreme mental and physical demands as well as traumatic situations [8]. Currently, declines in mental health is often recognized too late in time to intervene preventively.

While there is no recognized objective or continuous measurement of mental health yet, affective state monitoring may be of assistance in gaining insight into someone's mental health. Affective states shape the way that a person feels, thinks and behaves [9] and influence one's mental health in the long run, depending on their type, duration, intensity or frequency

A.-M. Brouwer is with The Netherlands Organization for Applied Scientific Research (TNO), Kampweg 55, 3769 DE Soesterberg, The Netherlands and with The Donders Institute for Brain, Cognition and Behavior, Thomas van Aquinostraat 4, 6525 GD Nijmegen, The Netherlands (e-mail: anne-marie.brouwer@tno.nl).

M. Schadd is with The Netherlands Organization for Applied Scientific Research (TNO), Oude Waalsdorperweg 63, 2597 AK The Hague, The Netherlands (e-mail: maarten.schadd@tno.nl)

H. Oldenhuis is with the Research Group Digital Transformation, Hanze University of Applied Sciences, Zernikeplein 11, 9747 AS Groningen, The Netherlands (e-mail: h.k.e.oldenhuis@pl.hanze.nl).

R. Verdaasdonk is with the TechMed Center, Faculty of Science and Technology, University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands (e-mail: r.m.verdaasdonk@utwente.nl).

L.van Gemert-Pijnen is with the Department of Psychology, Health and Technology, Centre for eHealth Research and Wellbeing, University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands (e-mail: j.vangemert-pijnen@utwente.nl).

TAFFC-2024-04-0220

[10], [11]. Affective state monitoring often relies on subjective measurements at sporadic time points [12]. Those subjective measures require effort from a person and commonly involve ratings of valence (pleasure/displeasure) and arousal (activation/deactivation) to infer one's affective state [13]. However, solely relying on those subjective measurements may not be reliable due to social biases (see [14]), recall biases [15], [16] or that individuals may not be completely honest about their actual affective state or are not fully aware of it [17]. Additionally, affective state measures in research have often been divided into discrete categories [18], [19], [20], [21] and thus may not reflect the complexity or level of the affective state [22].

A large amount of research on the relation between physical and affective states has been performed with wearables or smartphones [23], [24], [25], [26]. However, most of that research was limited in terms of not accounting for person-specific differences [27], [28], [29], the number and types of modalities measured [30], [31], the scale and/or duration [31], [32], the lack of ecological validity due to laboratory settings [33]; see [34]. These circumstances make it difficult to develop models that can describe an individual's state accurately and to generalize results to everyday settings.

Accordingly, predictions of changes in affective states at a subject-independent level showed only limited success, with small to moderate effects [35]. One reason for this may be that subject-independent prediction models do not take into account the large variability between individuals. In fact, the call for tailoring or personalization has existed already for a longer period of time (see [34]). Some research has shown the success of subject-dependent models for classification [36] and detection [32] of affective states, but were still often of short duration.

While previous research often employed a data-driven approach or solutions, Adler et al. [37] argued that the implementation needs more norms focusing on individuals' needs and protections, thus avoiding to collect all kinds of data that might be of interest. Also, Mohr, Shilton and Hotopf [38] argue for more personal sensing, highlighting a more sensitive form of surveillance since the collected data in real life from devices is intimate and personal. Such data and data-sharing can have different consequences for individuals, such as affecting life insurance premiums [39], highlighting the importance of explicit norms regarding appropriate data collection. One approach towards a more tailored data collection would be to involve and use the knowledge of a user to guide the process of which data needs to be collected. Intelligent systems such as human digital twins may be helpful when used for observing humans and their context [17], but the relationship of affective states with physiological and contextual measures is not fully understood yet.

Though previous research varied in terms of the sensors, signals and settings to make inferences about affective states, they all strove to find a more objective and continuous measure of mental health. Research on the detection of affective states with wearable sensors has shown promising results [10], [29], [40], [41], [42], [43], [44]. These results may also enable mental health detection in the long run since unhealthy affect is found in 40-75% of mental disorders [10], [40]. Previous research has

shown that affective states may be measured continuously and objectively, however, these results are often based on subject-independent models, thus leaving it unclear to what extent affective states can be predicted at an individual level.

In this study, we aim to investigate the relations between behavioral patterns, physiology and affective states in everyday settings. The contributions of this work are as follows:

- Collection and analysis of an extensive, longitudinal data set ($n = 16$ with approximately 299 days per participant), consisting of daily self-reported valence and arousal, combined with continuous physiological and behavioral sensing to create subject-dependent prediction models (individual-level).
- Empirical evidence that subject-dependent prediction models (individual-level) significantly outperform leave-one-subject-out subject-independent predictions (group-level) in this study, with varying performances indicating the importance of tailoring.
- Empirical evidence that subject-dependent prediction models (individual-level) outperform on average subject-independent prediction models (group-level) of previous studies such as [35]
- Preliminary evidence that self-evaluations of behavioral patterns can assist in the selection of the types of data before data collection takes place.

II. BACKGROUND AND RELATED WORK

A. Behavioral patterns and affective states

Through consumer-grade wearables and smartphones, self-reported behavioral patterns can be replaced by more continuous and objective measures, such as step count, GPS locations or trends in smartphone usage [2]. In affective state research, such sensors are used to obtain different types of observations and are then linked with self-reports of psychological concepts (e.g. stress, depression) serving as the ground truth [17]. In the following, the current state and shortcomings of stress and depression predictions with sensors are described.

In terms of stress, it has been shown that certain physiological and behavioral signals such as sleep quality, heart rate, heart rate variability or step count are indicative of perceived stress [45], [46], [47], [48], [49]. These signals can be passively monitored with sensors and wearables in real life. However, studies have shown mixed results in terms of stress prediction [42], [50], [51], [52]. The state-of-the-art is fairly accurate for predicting stress in laboratory settings, while real-life studies tend to perform worse than that [53], [54]. In real-life settings, errors occur more frequently due to technical errors, influences of context and activities, incorrect placement of wearables or detached equipment, which likely increases with multi-device measurements [44].

In terms of depression, it has been shown that people diagnosed with depression tend to isolate themselves from social situations [55], [56] experience sleep disturbances [57] or show lower levels of movement duration and intensity [58], [59], [60]. Research by [61] has also demonstrated significant differences between depressed and non-depressed groups in terms of GPS, phone usage, physical activity, sleep and mood.

Different types of studies have shown that depression symptoms could be predicted with phone usage data [26], [61], [62]. However, little is known about the development of mental health problems within the general population, including identifying and predicting problems to intervene preventively. Next, three research studies [35], [63], [64] are described due to their close resemblance with the current study.

Recent work of [35] used a multimodal approach by combining wearable and contextual sensing to predict daily ordinal stress levels rated on a 5-point Likert scale. In this study [35], approximately 47 days of data per participant were collected (28,226 samples in total from 597 participants). The passively sensed data consisted of physiological measurements, relative locations, smartphone usage and changes in environmental conditions [35]. The standard Random forest (RF), employing subject-independent folds and nested cross-validations, resulted in a Spearman's rho correlation between passively sensed data and self-reported stress levels of 0.19. When accounting for individual differences in stress sensitivity with the STAI questionnaire, the Spearman's rho correlation increased to 0.25. For future work, the authors suggested increasing the temporal resolution of stress levels, including more passive contextual information, applying time series modeling and exploring individual differences [35].

Within a non-clinical population, [63] used a multimodal approach to classify and predict depression status (depressed versus non-depressed) based on daily mood, smartphone and wearable data. The study included 54 participants and approximately 28 days of data per participant [63]. Daily aggregated features for comparison across the two groups included mood, sleep, physical activity, phone usage and GPS mobility. Mood was measured three times during the day, with valence and arousal rated separately on a 9-point Likert scale from -4 to 4 (low to high). For future work, the authors suggested using a clinical population, exploring both subject-dependent and subject-independent models, comparing subgroups and exploring differences across subjects in digital biomarkers [63].

Another study by [64] created multilevel models of depression and anxiety symptoms based on wearable and smartphone data. Sixty participants took part in a two-week study where data on their location, smartphone usage, activity, sleep and heart rate variability were passively collected [64]. Additionally, participants reported their daily mood three times a day, with valence and arousal rated separately on a 9-point scale from -4 to 4 (low to high). For future work, the authors suggested replicating the findings with a larger sample size and a clinical population, including Android users, broadening the sample in terms of demographics, and exploring smartphone app usage and individual differences in symptomatology [64].

In sum, previous studies such as [35], [63], [64] ranged in duration from 14 to 47 days per participant, which may not give enough insights into stable or deviating behavioral patterns of individuals. Predictions of stress and depression on a subject-independent level showed small-to-moderate effects and in accordance with future work recommendations, the current work focuses on subject-dependent prediction models.

B. Classification types of affective states

Previous research sometimes focuses on the prediction of solely negative concepts such as depression or stress in a more categorical or ordinal manner [35], [63], [64]. One issue with those discrete distinctions or categorizations is that they do not necessarily reflect the whole spectrum of states since affective states, such as feeling stressed or depressed, can also lie somewhere in between the categories or distinctions [22].

The research of [22] took the first step toward predicting valence and arousal on a continuous scale in a laboratory setting. The WESAD dataset [65] was used for the analysis, containing data about acceleration, skin temperature, electrodermal activity, heart rate, blood volume pulse and heart rate variability. It was shown that Long Short-Term Memory (LSTM) models could reliably predict fine-grained levels of valence based on wrist-worn sensors [22].

In sum, previous research commonly made more concrete categorizations of affective states [35], [63], [64], such as binary, three- or four-class. Recent research [22] also looked into more fine-grained valence and arousal classifications. The current work also associates with a more fine-grained classification since affective states are more complex than discrete binary or categorical classifications.

C. Subject-dependent versus subject-independent models

As mentioned before, current research focusing on subject-independent predictions has shown limited success with small-to-moderate effects (e.g. [35]). One possible reason for this is that subject-independent models do not take into account variations across individuals. More specifically, individuals tend to differ in their affective and physical experience as well as possible reasons for threats to mental health [15], [66], [67]. Furthermore, individuals also tend to differ in terms of stressors, with some perceiving some tasks or variety of work as easy and others as difficult or stressful [17].

Finally, individuals also differ in their coping responses and manifestations. For instance individuals tend to differ in how stress changes their eating behavior [68]. For some individuals, food intake may increase, decrease or not change at all. Similarly, while one person might isolate themselves in stressful situations, another person might search for social contact to help alleviate the feelings of stress.

In sum, both perceptions and manifestations of affective states, visible in terms of behavioral and physiological data, are highly person-dependent, which is why one should avoid a "one-size-fits-all" approach [17], [67]. In support of this idea, [22] has shown that the performance of subject-dependent prediction models differ across subjects and that using the same set of features and sensors for each individual is not necessarily desirable. Other research has shown notably higher accuracies for subject-dependent prediction models (see [17]). However, it is worth noting that subject-dependent prediction models only tend to be more accurate if a sufficient number of self-reported stress labels are provided to train the model (see [17]).

D. Novelty of current study

This study aims to work towards a better prediction of affective states at an individual level. To do so, a fine-grained prediction of affective states with continuous measures of

valence and arousal will be applied to get a complete picture of the person [22]. In contrast to a laboratory setting in [22], the current study aims at predicting fine-grained (day-to-day) levels of valence and arousal in daily life. As mentioned in the previous paragraphs, previous studies are often of short duration [35], [63], [64]. The current data collection consists of a period of nine months per participant with multimodal data, combining subjective data with physiological and behavioral patterns. Physiological patterns are recorded and explored using wearables, focusing primarily on physical activity and sleep. Behavioral patterns are recorded and explored with smartphone sensors and thus include smartphone usage, GPS motion and communication. These physiological and behavioral patterns of a day are used to predict an individual's self-reported affective state of that particular day. The longitudinal design aims at capturing a sufficient variety of affective states and eventually providing more accurate and representative data [69], [70].

Preliminary results of the current work have shown small-to-moderate correlations between predictors and daily measurements of valence and arousal at a subject-independent level [71]. Additionally, the standard deviations of correlations gave a first indication of possible differences between individuals, thus suggesting the merit of an individual-based approach for prediction models. In accordance with this, recent research has shown that subject-dependent models tend to outperform leave-one-subject-out (LOSO) models [72] which will also be tested in the current study. Thus, it is expected that leaving one subject out and predicting values of that subject with the remaining subjects of the sample will perform worse than predictions on a subject-dependent level.

Lastly, the current study also aims to gather additional subjective input from participants to triangulate physiological, behavioral, and affective state data with a priori knowledge to better understand that person's mental health. With more knowledge about the individual and their behavioral patterns, it might become possible to develop norms for an appropriate data collection by tailoring the types of data that need to be collected for an individual.

Based on previous work, the current study will test the following hypothesis:

- 1. Subject-dependent models will outperform subject-independent models (leave-one-subject-out) in predicting affective states (valence and arousal) of individuals using passively sensed data

Furthermore, we will explore the following...

- 2. What are the most important features for subject-dependent models of *valence* levels and to what extent do these features differ across subject-dependent models?
- 3. What are the most important features for subject-dependent models of *arousal* levels and to what extent do these features differ across subject-dependent models?
- 4. To what extent does a priori self-evaluations of the relation between behavioral patterns and not feeling mentally well correspond with subject-dependent models of valence?

III. METHODS

The research proposal “CAMSTAM – Continuous Automatic Mental State Monitoring” was reviewed and accepted by the TNO Institutional Review Board (IRV), number: 2021-078. The preliminary study protocol can be found in [71].

A. Participants

This study recruited PhD candidates as study participants. PhD candidates are a risk group due to the high workload and cognitive demands, with one in two PhD students experiencing psychological distress and one in three running a higher risk of developing a common psychiatric disorder [73]. Sixteen PhD candidates who are currently working in the Netherlands took part in a nine-month-long study starting at the end of March 2022. Upon reaction on a call for participation, interested applicants were invited for an in-person intake interview with the main researcher of the study. Participants that were included in the study needed to be working as a PhD candidate, interested in the use of wearables and self-monitoring, experiencing no heart-rhythm disorder, willing to deliver sufficient data within nine months of the data collection period and having a sufficient level of Dutch proficiency to fill in the questionnaires. Different types of questionnaires were filled in during the intake and on a monthly basis. These questionnaires were excluded in the current study as the focus is on day-to-day variability. For compatibility reasons with the AWARE framework [74], used to collect behavioral data, participants were also required to own an Android smartphone. Additionally, participants received a smart ring from Oura [75] for the collection of physiological data (see description Section III.C.). The Oura ring was shown to be a reliable instrument for physiological measures including wake-sleep detection [76], Sleep Onset Latency (SOL) Total Sleep Time (TST) and Wake after Sleep Onset (WASO) [77], [78], step count, active and total energy expenditure [79]. Previous research on physiological monitoring in daily life for predictions of mental health have used the Oura ring as well [64], [80], [81].

The age of participants ranged from 24 to 38 years, with a mean age of 30 years ($SD = 4.02$). Nine (56,25%) of the participants were female, and 7 (43,75%) were male. At the beginning of the experiment, 2 (12,5%) were in their first year of the PhD trajectory, 6 (37,5%) in their second year, 3 (18,75%) in their third year and 5 (31,25%) in the last year. As a compensation for their efforts in this relatively intensive data collection, participants received €10 each month for their participation and also had individual monthly feedback sessions with the main researcher to get insights into their raw data. To ensure a sufficient and continuous number of affective state reports and physiological measurements for the predictive modeling, we required 85% of valid data points (days with both subjective and physiological data present) which equals roughly 230 data points or 7.6 months of data collection. To facilitate participant recruitment and optimize adherence during the participation, participants who delivered these 85% of valid data points were rewarded with keeping the Oura ring that they used for the data collection. If participants did not deliver 85% of valid data points within the nine months, the data collection was extended to a maximum of eleven months. In four cases,

the 85% could not be reached due to connectivity issues with the Oura ring or technical difficulties with the Ecological Momentary Assessment (EMA) prompts towards the end of the study. These participants were still included in the analyses.

B. Affective State Assessment

At the end of each day, from 5 p.m. until 5 a.m. of the following day, participants filled out two questions about their affective state during the day. The “How am I” Ecological Momentary Assessment (EMA) application developed by TNO [82] was used for those two EMA questions and also for prompting participants to fill in the questionnaires. To ensure well-timed prompts, participants could select the timing of the prompts themselves. For a fine-grained measurement of valence and arousal, the Circumplex Model of Affect by [13] was used as a basis. During the first question, participants rated on a scale from 0 (unpleasant) to 100 (pleasant) the valence level of their day. Secondly, participants rated on a scale from 0 (mentally passive) to 100 (mentally active) the arousal level of that day. Similar low-burden EMA measurements of valence and arousal were also recently used by [83], [84].

TABLE 1
FEATURES EXTRACTED FROM THE OURA RING AND AWARE FRAMEWORK

Sensor	Signals	Features
Oura Ring	Motion, heart pulse	Sleep start time, Sleep end time, SOL, TST, SE, Fluctuations in TST, Fluctuations in SE, Mean nocturnal heart rate, Lowest nocturnal heart rate, Time to lowest nocturnal heart rate, Nocturnal HRV, HRV during last 30 minutes of sleep, Fluctuations in nocturnal HRV, Sedentary time, MVPA, MET-minutes
AWARE framework	Phone usage, GPS, motion	Screen time, Browser app time, Communication app time, Entertainment app time, Gaming app time, Information app time, Social app time, Commerce app time, Total app time, Unique apps per day, Notifications per day, Social notifications per day, Unique social contacts per day, Time at home, Time at work, Time at sports location, Time at other known location, Time spent traveling, Distance traveled, SMS messages sent, Number of incoming phone calls per day, Number of outgoing phone calls other per day, Duration of incoming phone calls, Duration of outgoing phone calls

Note. SOL = Sleep Onset Latency; TST = Total Sleep time; SE = Standard Error; HRV = Heart Rate Variability; MET = Metabolic Equivalent of Task; MVPA = Moderate-to-Vigorous Physical Activity

C. Sensing Devices and Predictors of Affective States

Participants used two sensing devices to deliver physiological, behavioral and contextual data. An overview of the data collected can be found in Table 1. With the commercial-grade smart ring from Oura [75] participants provided data about their daily activity and sleep patterns. Such measures have been previously linked to affective states [48], [49], [63], [85], [86], [87], [88], [89], [90]. Most of the Oura features could be immediately exported via the Oura dashboard. To ensure sufficient Oura data per day, the non-wear time of subjects was closely monitored. If the non-wear time increased, the corresponding subject was contacted via e-mail. An overview of the selected Oura features can be found in Table 1. See section III.E. for an overview of how the features are defined and how they are measured.

Furthermore, participants installed a customized smartphone application (created with the AWARE framework, see [74]) to deliver smartphone usage metrics such as number of phone unlocks, application usage and GPS locations. Smartphone usage has been linked to affective states as well, including aspects such as phone unlock duration, notifications and social media [91], [92], [93], [94]. For the current study, interactions with the smartphone were measured by screen-on durations, number, types and duration of apps, number, type and duration of phone calls and sent messages. Staying at a certain location and deviations from location patterns were also found to be connected to affective states [25], [63], [95], [96], [97]. The raw GPS data, with input about location context from individuals, was used to create location-based features. These included time at work, home, sports, social and unknown locations, total distance traveled, and time spent traveling. For the AWARE app, raw data about screen time could be immediately exported, but the rest of the features had to be pre-processed and transformed into features (see section III.E. for description). To create daily values of each feature, the timeframe of a day had to be determined for each individual separately to account for individual behavior, such as filling in a questionnaire after midnight for the day before, or being on vacation in a different timezone.

D. Data Collection Protocol

In total, the data collection process took eleven months to finish, targeting at 276 valid data points per participant (see Section III.A.). First, participants provided written informed consent for the different types of data that were going to be collected. Then, participants were asked in an intake interview about the GPS coordinates of their most-visited locations, daily routines and behavioral patterns when they felt stressed or did not feel well. To explore whether a participant’s self-knowledge can lead to valid feature selections before the data collection, participants were asked to describe their (self-estimated) behavioral patterns when feeling stressed or down. This was done for each of the following categories: exercise/physical activity, sleep, movement/locations, app usage, social contacts, smartphone usage and communication via smartphone such as calling and SMS. These types of categories were previously shown to be linked to affective states (see Section III.C.) and will therefore be tested and compared in both a subjective and objective manner.

TAFFC-2024-04-0220

Additionally, a custom-sized Oura ring was ordered for each participant. The ring size and color were tailored to each individual to optimize measurement accuracy and participant adherence. Participants were instructed to wear the ring as much as possible, excluding situations such as charging the device or when they were doing sports where the ring could hinder movement or be damaged. To motivate participants and ensure their adherence throughout the nine months, the main researcher was regularly in contact with each participant and planned a feedback moment of 30 minutes every four weeks. During their individual feedback moment, participants got insight into their own personal raw data, could exchange experiences with the researcher or report problems with the data collection. After delivering at least 85% valid data points (see Section III.A.) which equals approximately 274 days, an outtake interview was planned. During the outtake interview, participants were asked to evaluate and give feedback about their participation in the research.

E. Data Preparation

For each predictor variable, daily values were aggregated and merged into one complete dataset. The data preparation will be explained in short here. A more elaborate description of the pre-processing and variables can be found in the supplemental materials (see Appendix A).

First, the raw AWARE data was extracted and pre-processed into the features mentioned in Table 1. This included removing noise, outliers and faulty data, such as crashes, the smartphone running out of battery or incorrect logging of app usage when the screen was turned off. Additionally, time stamps had to be aligned for participants that traveled in different time zones. For the locations, additional steps for pre-processing included translating lat-lon positions to location categories indicated by participants to calculate how much time they spent at each location. For smartphone usage, system apps were removed that were continuously running, technical notifications were removed that are not notifications for a user and used apps were translated into app categories. The number and count of calls were exported manually at the end of the data collection since most participants' tracking with the AWARE app did not function properly. From the call logs, missed calls and all calls with a duration of 10 seconds or less or two hours and more were removed.

Secondly, the Oura data was extracted from Oura on the web and pre-processing was performed in RStudio (version 2022.12.0 Build 353) using R (version 4.2.2). Most features mentioned in Table 1 could be exported directly. Features such as Sleep Onset Latency (SOL) were transformed from seconds to minutes and Total Sleep Time (TST) being transformed from seconds to hours. Lastly, a 7-day rolling coefficient of variation (CV) was calculated for TST and Sleep Efficiency (SE), reflecting intra-individual variations (based on [98]).

Thirdly, the daily valence and arousal levels were extracted from the How am I app. Participants could fill in the questions between 5 p.m. and 5 a.m. at a time of their own convenience. If two questionnaires were filled in on the same day, one of the two questionnaires was either removed or transformed. Removal took place if there was an error with the app or data transfer, for instance having two identical answers per question

in the dataset. Attributing to the previous day took place if questionnaires were filled in after midnight and before 5 a.m. of the next day or when being in a different time zone.

Depending on the type of feature, missing daily values were replaced with either a 0 or with the personalized means of each participant. For instance, for different app types such as Social Media daily values were replaced with 0 since it could be the case that participants did not use that type of app on that particular day. In contrast, for Oura features and locations such as Home, daily missing values were replaced with personalized means due to the high likelihood of participants being some part of the day at home, sleeping and moving. Each feature and outcome variable (valence, arousal) was baselined for each individual using z-score normalization. This involved subtracting the individual's mean value of all observations from each observation of that individual and dividing by the standard deviation of all observations of that individual, resulting in standardized values. A value of 0 represented the personal mean, and values between -1 and +1 indicated one standard deviation above or below the personal mean of that test subject.

F. Data Analysis

The prepared data was analyzed using Python 3.10.8, kits Pandas 1.5.3, Scikitlearn 1.2.2. Statistical analyses were performed in RStudio [99]. Similarly to [72], Random forests (RF) [100] were used as the learning algorithm since i) the collected data is tabular data, with tree-based models being useful for tabular data [101], ii) RF are computational-wise fast, iii) RF require less parameter tuning and have better generalization [102], [103] and iv) RF is reliable in terms of skewed or scarce (outliers) training examples. In addition, similar prediction studies [35], [80] compared different classifiers and have shown that RF outperforms other classifiers. These aspects make RF useful for the prediction models of the current study.

RF were used for creating the prediction models of fine-grained valence and arousal. Subject-dependent models were trained on 80% of the individual's randomly sampled data and tested on the other 20% of the individual's data. The split of train and test set was randomized and repeated with 100 tests. The feature vector had 45 values, with each value representing one of the predictors mentioned in Table 1. In turn, the values of the predictor and outcome variables are daily values, with each value representing a different day of an individual. The predictors used to train the RF model were selected based on the feature importance of that individual. The feature importance was determined based on the increase in mean squared error (MSE) of predictions.

Furthermore, leave-one-subject-out (LOSO) subject-independent models were created to predict the valence and arousal levels of the left-out participant. The subject-independent LOSO prediction models were trained on (randomly selected) 80% of the data from 15 individuals and tested on (randomly selected) 50% of the data of the left-out individual. Both subject-dependent and subject-independent test-train sets were randomly split. Compared to the subject-dependent prediction models, the subject-independent LOSO prediction models contained more data points in the train and test set. Thus, the test sets per run were not equal across subject-

TAFFC-2024-04-0220

dependent and subject-independent LOSO models. By running 100 tests with random splits, the differences that could arise due to the random splitting will be averaged out.

The value of the tree-depth hyperparameter also differed across subject-dependent and subject-independent models, based on preliminary analyses. More specifically, tree-depth 8 was used for the subject-independent prediction models as a large amount of data is available. For the subject-dependent models, a tree depth of 3 was chosen to avoid overfitting. With the continuous values of valence and arousal, we could not apply metrics such as F1 or Kappa that are only suitable for binary and/or categorical labels [104]. Following [35], the Spearman's rho correlation coefficient is derived from the RF and describes the correlation between the observed and the predicted scores, capturing the continuous nature of the arousal and valence scores. In this context, it captures the extent to which relative differences in valence and arousal levels on different days are (rank-order) correlated with the prediction (see [35]). The Spearman's rho correlation coefficient was also used to compare the performance of subject-dependent and leave-one-subject-out subject-independent prediction models.

For the hypothesis that 'subject-dependent models will outperform subject-independent (leave-one-subject-out) models in predicting affective states (valence and arousal) of individuals using passively sensed data', two one-sided paired sample t-tests were conducted to evaluate whether there was a difference between subject-dependent and subject-independent LOSO models, with a significance level of 0.05. With the first paired sample t-test, Spearman's rho values of subject-dependent and subject-independent prediction models of valence were compared. The second paired sample t-test compared Spearman's rho of subject-dependent and subject-independent prediction models of arousal.

To explore the second and third research question 'What are the most important features for subject-dependent models of valence/arousal levels, and to what extent do these features differ across subject-dependent models?', RF were run multiple times on the subject-dependent prediction models. A sequential feature selection with backward elimination was used to find the most important features per model [105]. More specifically, the analysis first included all prediction features, and then every time the least important feature of the model (increase in mean squared error MSE) was removed, progressing to a model with only one feature. The feature selection mechanism was based on the change in performance of a RandomForestRegressor after iteratively dropping a feature. To find robust feature importance values in the experiment, the RandomForestRegressor was trained 100 times on the dataset and each time 20% of the data points were left out. This approach minimizes the influence of chance, as it is very unlikely for a suboptimal feature to appear advantageous 100 times. Preliminary experiments of the dataset showed that results might vary slightly when rerunning the same experiment. These small differences can be observed at two decimal places or more. Such variability might cause a feature to be discarded slightly earlier or later in the iteration process, but the general value of a feature remains consistent.

In comparison to an additive approach, the iterative feature removal also takes interdependencies between variables more

into account [106]. A similar approach of iterative feature removal was used by [107] and also showed best classification power in [106]. The model with the highest Spearman's rho correlation coefficient was chosen as an indicator to decide which and how many features should be selected for each individual. For model exploration, the features of subject-dependent models of valence and arousal were compared across individuals to test for overlap.

To explore the fourth research question 'To what extent does a priori self-evaluations of the relation between behavioral patterns and not feeling mentally well correspond with subject-dependent models of valence?', self-indicated behavioral patterns when participants did not feel mentally well were compared to the most important features in the models of the individual participants. For this part of the analysis, the number of most important features was based on the mean squared MSE. The cut-off point was determined when the feature importance reached 0 or became negative, indicating no added value or a negative influence on the model performance. First, the expected change was described in a table and translated into binary categories. When a participant expected a change (either positive or negative), it was categorized as "1" and when it was expected to be equal it was categorized as "0". Next, all existing features were put into the seven categories (see Appendix B). For instance, the features of MVPA, MET and sedentary time were categorized into 'physical activity' and time spent at a friend's or family's place (social locations) was categorized into social contacts. Lastly, it was investigated whether the expected change of the participant was confirmed by the most important features of the associated subject-dependent models.

IV. RESULTS

A. Descriptives

In Figure 1, valence and arousal scores from all participants are depicted. In total, there were 4,782 observations with 4,220 (88.25%) valid observations of valence and arousal. A total of 562 (11.75%) observations missing either valence or arousal values – these were removed. On a participant level, the amount of valid data ranged from 237 (72.77%) days of valid observations for participant 8 to 281 (99.29%) days of valid observations for participant 7 (see Appendix C).

For the valence scale, 82.44% ($N = 3,479$) of all valid observations were higher than 50 ($M = 65.79$, $SD = 18.46$), indicating an overall tendency to report pleasant states. For the arousal scale, 73.58% of all valid observations ($N = 3,105$) were higher than 50 ($M = 61.94$, $SD = 20.48$), indicating an overall tendency for mentally active states. These tendencies are also visible in Figure 1, with the upper right quadrant of happiness being most prominent in comparison to the other three quadrants.

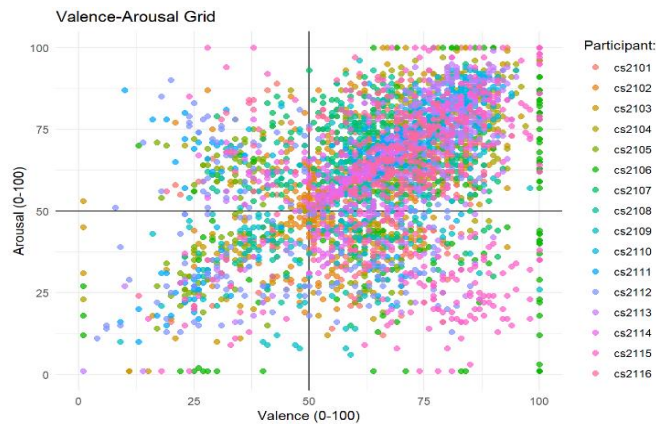


Fig. 1. Affective states of all participants distributed across quadrants

There are large differences in valence and arousal scores between participants. For example, 98.05% ($N = 252$) of all valid valence observations of participant 10 were higher than 50 ($M = 80.76$, $SD = 8.29$), see Figure 2. 98.44% ($N = 253$) of all valid arousal observations of this participant were higher than 50 ($M = 81.42$, $SD = 8.08$). In contrast, for participant 12 (see Figure 3), only 59.11% ($N = 146$) of all valid valence observations were higher than 50 ($M = 53.13$, $SD = 20.40$), and only 48.58% ($N = 120$) of all valid arousal observations were higher than 50 ($M = 49.26$).

To investigate relative changes within a participant, valence and arousal values were baselined. Baselining these for each individual shifted the center point for participant 10 to the upper right quadrant while it stayed almost the same for participant 12.

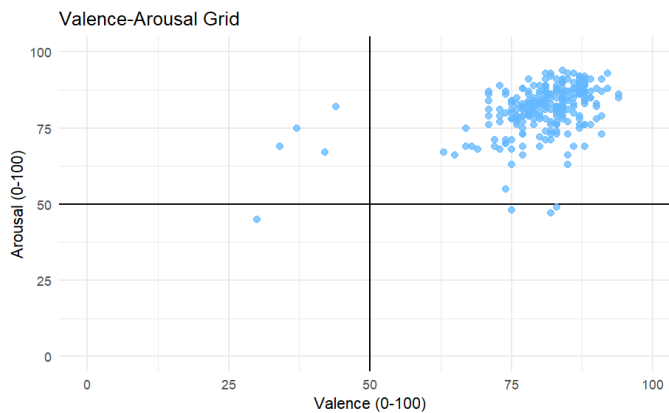


Fig. 2. Affective states of participant 10

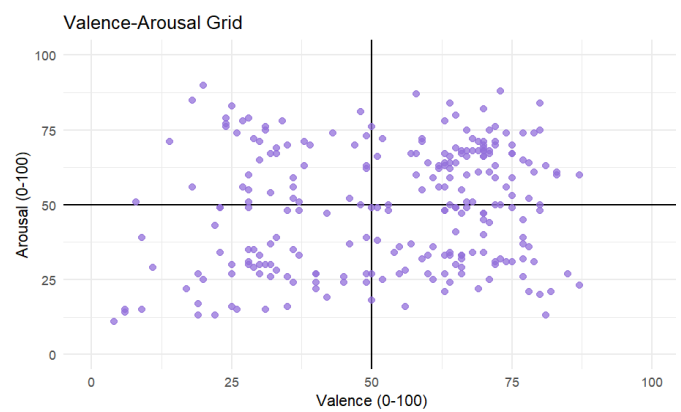


Fig. 3. Affective states of participant 12

In Table 2, descriptive statistics of the AWARE and Oura features can be found. For the AWARE measurements, there were 3,953 (82.66%) days of valid observations and 829 (17.34%) days of observations with missing values (see table Appendix C). On a participant level, these ranged from 180 (55.05%) days of valid observations for participant 8 to 281 (99.64%) days of valid observations for participant 6. For the Oura measurements, there were 4,290 (89.71%) valid observations and 429 (10.29%) observations with missing values (see Appendix C). On a participant level, these ranged from 208 (64.61%) days of valid observations for participant 8 to 275 (99.63%) days of valid observations for participant 4.

B. Prediction models of valence and arousal levels

First, Spearman's Rho values and standard deviations were calculated, averaged and compared for subject-dependent and LOSO models across 100 runs (see Table 3). The robustness of subject-dependent predictions was established by contrasting the LOSO performance with the subject-dependent prediction for the individual left out of the LOSO model.

As can be seen in Table 3, subject-dependent prediction models of valence and arousal widely vary both in terms of Spearman's Rho values as well as number of selected features based on the highest Spearman's Rho correlation coefficient. The averaged Spearman's Rho correlation across all subject-dependent models for predicted versus observed valence levels was 0.30. For LOSO models of valence, the averaged Spearman's Rho correlation was 0.17. For arousal levels, the averaged Spearman's Rho correlation of subject-dependent models was 0.36. The LOSO models of arousal had an average Spearman's Rho correlation of 0.28.

TABLE 2

SUBJECT-INDEPENDENT AVERAGE AND SUBJECT-DEPENDENT RANGES PER FEATURES

Feature	M (SD)	Min M (SD)	Max M (SD)
Screen time (mins)	191.8 (103.7)	87.3 (41.0)	324.0 (122.8)
Communication app time (mins)	16.8 (21.2)	5.3 (11.1)	37.4 (34.6)
Social Media app time (mins)	16.1 (25.1)	0 (0)	46.1 (36.0)
Duration incoming calls (mins)	2.1 (7.1)	0.3 (1.5)	7.6 (17.4)
Duration outgoing calls (mins)	3.0 (10.0)	0.5 (2.3)	10.2 (19.8)
Time at home (mins)	503.1 (460.9)	0 (0)	1075.7 (348.1)
Time at work (mins)	135.5 (181.5)	0 (0)	313.2 (268.0)
Time at social locations (mins)	102.4 (275.8)	0 (0)	486.0 (395.1)
Distance traveled (km)	112.6 (283.0)	2.96 (11.1)	1094.3 (304.8)
Time spent travelling (mins)	40.6 (63.7)	2.9 (8.4)	182.9 (26.5)
TST (mins)	7.2 (1.1)	6.0 (1.1)	7.8 (1.2)
HRV	4.1 (0.5)	3.2 (0.3)	4.9 (0.2)
SOL (mins)	10.5 (9.7)	4.9 (3.7)	16.4 (12.0)
Mean nocturnal heart rate	58.1 (7.8)	43.5 (3.1)	67.1 (4.2)
Lowest nocturnal heart rate	51.1 (7.0)	39.2 (2.6)	59.5 (5.9)
Sedentary time (mins)	500.9 (134.2)	444.3 (94.9)	566.3 (83.6)
MVPA (mins)	73.7 (58.1)	29.5 (19.6)	140.4 (83.3)
MET (mins)	513.7 (335.9)	305.9 (123.7)	775.0 (345.1)

Note. SOL = Sleep Onset Latency; TST = Total Sleep time; SE = Standard Error; HRV = Heart Rate Variability; MET = Metabolic Equivalent of Task; MVPA = Moderate-to-Vigorous Physical Activity

The hypothesis posited that ‘subject-dependent models will outperform subject-independent (leave-one-subject-out) models in predicting affective states (valence and arousal) of individuals using passively sensed data’. Overall, for 15 out of 16 participants (93.75%), subject-dependent prediction models of valence with feature selection outperformed the corresponding subject-independent LOSO prediction model. For one participant (participant 5), the subject-dependent prediction model of valence performed worse than the subject-independent LOSO prediction model of valence. The results of the first paired sample t-test indicated that subject-dependent models of valence ($M = 0.30$, $SD = 0.11$) performed significantly better than the subject-independent LOSO models ($M = 0.17$, $SD = 0.06$; $t(15) = 5.30$, $p < 0.001$).

TABLE 3

SPEARMAN’S RHO CORRELATIONS OF LEAVE-ONE-SUBJECT-OUT (LOSO) SUBJECT-INDEPENDENT VERSUS SUBJECT-DEPENDENT PREDICTION MODELS WITH VALENCE AND AROUSAL

P	ρ valence subject-dependent (σ ; N features)	ρ valence LOSO subject-independent (σ)	ρ arousal subject-dependent (σ ; N features)	ρ arousal LOSO subject-independent (σ)
1	0.31 (0.10; 2)	0.24 (0.06)	0.69 (0.07; 3)	0.6 (0.04)
2	0.20 (0.11; 2)	0.14 (0.07)	0.26 (0.12; 1)	0.05 (0.06)
3	0.25 (0.14; 9)	0.18 (0.06)	0.43 (0.10; 5)	0.26 (0.06)
4	0.37 (0.12; 6)	0.17 (0.06)	0.55 (0.10; 6)	0.48 (0.05)
5	0.14 (0.11; 1)	0.17 (0.07)	0.23 (0.11; 3)	0.24 (0.06)
6	0.23 (0.11; 2)	0.13 (0.07)	0.27 (0.10; 3)	0.24 (0.06)
7	0.60 (0.07; 6)	0.38 (0.06)	0.33 (0.11; 2)	0.20 (0.06)
8	0.25 (0.13; 2)	0.18 (0.07)	0.25 (0.13; 5)	0.26 (0.07)
9	0.43 (0.09; 6)	0.14 (0.05)	0.59 (0.09; 3)	0.56 (0.04)
10	0.33 (0.13; 8)	0.11 (0.06)	0.33 (0.12; 4)	0.32 (0.06)
11	0.25 (0.11; 4)	0.09 (0.07)	0.29 (0.11; 5)	0.24 (0.06)
12	0.40 (0.11; 3)	0.02 (0.07)	0.35 (0.12; 4)	0.25 (0.07)
13	0.22 (0.13; 2)	0.10 (0.06)	0.24 (0.12; 17)	0.06 (0.06)
14	0.36 (0.11; 4)	0.21 (0.06)	0.27 (0.12; 5)	0.08 (0.07)
15	0.14 (0.12; 7)	0.08 (0.07)	0.53 (0.09; 2)	0.49 (0.05)
16	0.38 (0.11; 9)	0.32 (0.07)	0.16 (0.12; 6)	0.16 (0.07)
M	0.30 (0.11)	0.17 (0.06)	0.36 (0.11)	0.28 (0.06)

Note. P = Participant; ρ = Spearman’s rho averaged over 100 runs; M = mean across all participants; σ = standard deviation of 100 runs

Overall, for 13 out of 16 participants (81.25%), subject-dependent prediction models of arousal with feature selection outperformed the accompanying subject-independent prediction model. For one participant (participant 16), the subject-dependent prediction model of arousal performed equally well as the subject-independent LOSO prediction model of arousal. For another two participants (participants 5, 8), the subject-dependent prediction model of arousal performed worse than the subject-independent LOSO prediction model of arousal. The results of the second paired sample t-test indicated that subject-dependent models of arousal ($M = 0.36$, $SD = 0.11$) performed significantly better than the

TAFFC-2024-04-0220

subject-independent LOSO models ($M = 0.28$, $SD = 0.06$; $t(15) = 4.25$, $p < 0.001$).

Both for valence and for arousal, the predictions were significantly higher for subject-dependent models in comparison to subject-independent LOSO models, thus indicating support for the hypothesis.

C. Exploration of subject-dependent models

Regarding the second research question ‘What are the most important features for subject-dependent models of valence levels, and to what extent do these features differ across subject-dependent models?’, subject-dependent models of valence were explored per person based on an individual’s most important features and compared in terms of overlapping features across individuals. The number of features was determined for the best performing subject-dependent prediction model, as determined by the highest Spearman correlation coefficient that could be achieved. At a subject-dependent level, Spearman correlations for predicted versus observed valence ranged from 0.12 for participant 15 to 0.60 for participant 7 (see Table 3). The number of important features ranged from 1 for participant 5 to 9 for participant 3 and participant 16. For subject-dependent prediction models of valence, most overlap between individuals was observed in Oura measurements (see Appendix D). The highest overlap was for the hour when a person is waking up (sleep end hour; 7 out of 16 participants), followed by the time spent on sedentary activities (sedentary time; 5 out of 16 participants) and the covariate of a person’s total sleep time (Sleep TST-CV; 5 out of 16 participants). Additionally, the relation between the most important predictive features and valence levels differed both in terms of strength and direction across participants.

To explore the third research question ‘What are the most important features for subject-dependent models of arousal levels, and to what extent do these features differ across subject-dependent models?’, subject-dependent models of arousal were explored per person based on an individual’s most important features and compared in terms of overlapping features across individuals. At a subject-dependent level, the Spearman correlations for arousal ranged from 0.16 for participant 2 to 0.70 for participant 1 (see Table 3). The number of important features ranged from 1 for participant 2 to 17 for participant 13. For subject-dependent prediction models of arousal, the type of features with most overlap between participants was higher than valence in terms of count but slightly more mixed in terms of feature type, with the first two from Oura measurements and the third one from the AWARE app (see Appendix E). The highest overlap was for a person’s metabolic equivalent of task (MET) minutes (10 out of 16 participants), followed by the time spent on sedentary activities (sedentary time; 9 out of 16 participants) and the time spent on social media apps (Social Media; 6 out of 16 participants). Additionally, the relation between the most important predictive features and arousal levels differed both in terms of strength and direction across participants.

D. Overlap self-evaluation and collected data

To explore the fourth research question ‘To what extent does a priori self-evaluations of the relation between behavioral patterns and not feeling mentally well correspond with subject-dependent models of valence?’, self-reported knowledge about behavioral patterns related to not doing mentally well was visualized in Table 4 and compared to the results of the associated subject-dependent model (Table 5).

TABLE 4
BEHAVIORAL PATTERNS WHEN DOING NOT WELL MENTALLY ACCORDING TO PARTICIPANTS

P	[1] PA	[2] S	[3] L/C	[4] AU	[5] SC	[6] SU	[7] CS
1	-	+	-		-		
2	+	+		+		+	
3	+	-	+	+	+	+	-
4							
5	+		-	+	+	+	+
6	-	+	-	+	-	+	+
7	-	+	-	+	-	+	-
8	-			+		+	
9	-	+		+	-		-
10	-			+			
11	-	-	-	-	-	-	-
12	-	-			-		-
13	+	-	-		+		
14	+	+		+	-	+	-
15	-	+			-		
16	-	-	-	+	-		+

*Note. + = longer, more, better; - = worse, less, different times; empty space = more or less equal; P = Participant; PA = Physical activity; S = Sleep; L/C = Location/commutes; AU = App usage; SC = Social contacts; SU = Smartphone usage; CS = Communication via smartphone

Participants’ scores are represented based on their intake interview in which they indicated how their behavioral patterns change (or do not change) when they are mentally not doing well (see Table 4). The responses in the interview were categorized into seven categories, namely physical activity, sleep, locations/commutes, social contacts, smartphone usage and communication via smartphone. In the dataset, baselined valence levels ranged from unpleasant to pleasant, with not doing mentally well defined as the unpleasant levels of valence. To compare a priori knowledge with subject-dependent models of valence, the most important features of each subject-dependent model were put into the seven categories (see Appendix B) and then compared with whether a change was expected a priori (1) or not (0). For example, if MVPA was an important feature and the individual indicated an expected change in physical activity, this was categorized as a ‘hit’ (see Table 5).

In contrast to the previous analysis, the number of features were selected based on their added value (MSE) and not based on the Spearman’s rho correlation. Features that received zero or negative weight were excluded for this analysis, ranging from 38 excluded features out of 45 features for participant 8 and 11 to 24 excluded features out of 45 features for participant

7 (see last row of Table 5).

TABLE 5
RELATIONS EXPECTED BY THE PARTICIPANT AND FOUND RELATIONS FOR VALENCE LEVELS

P	[1] PA	[2] S	[3] L/C	[4] AU	[5] SC	[6] SU	[7] CS	T	%	N FI
1	hit	hit	hit	miss	miss	hit	hit	6/7	86	9
2	hit	miss	hit	miss	miss	hit	miss	4/7	57	14
3	hit	hit	hit	miss	miss	hit	hit	6/7	86	16
4	miss	miss	miss	miss	miss	miss	miss	2/7	29	12
5	hit	miss	hit	miss	miss	hit	hit	4/7	57	9
6	hit	hit	hit	miss	miss	hit	hit	6/7	86	16
7	hit	hit	hit	hit	hit	hit	hit	7/7	100	21
8	hit	miss	hit	miss	miss	hit	hit	6/7	86	7
9	hit	miss	miss	miss	miss	hit	miss	4/7	57	15
10	hit	miss	hit	miss	miss	hit	miss	5/7	71	10
11	hit	miss	miss	miss	miss	miss	miss	3/7	43	7
12	hit	miss	miss	miss	miss	miss	miss	4/7	57	11
13	hit	miss	miss	miss	miss	miss	miss	4/7	57	12
14	hit	miss	miss	miss	miss	miss	miss	4/7	43	10
15	hit	miss	miss	miss	miss	miss	miss	4/7	57	11
16	hit	miss	miss	miss	miss	miss	miss	5/7	71	17

*Note: **hit**: relation expected and confirmed by the data; **false positive**: relation expected but not confirmed by data; **miss**: no relation expected but relation found in data; **correct rejection**: no relation expected and confirmed by data; T = total agreement of categories; % = T translated into percentage; N FI = number of features based on feature importance; P = Participant; PA = Physical activity; S = Sleep; L/C = Location/commutes; AU = App usage; SC = Social contacts; SU = Smartphone usage; CS = Communication via smartphone

Table 5 shows the overlap between the categories of the intake interview and the categories of the collected data. The numbers in the left column of Table 5 represent the categories of behavioral patterns as indicated in the left column of Table 4. For 7 out of 16 participants (43.75%), the overlap was at least 71%, for 6 participants (37.5%) it was 57% and for 3 (18.75%) it was lower or equal to 43%. The fifth category, social contact, had the lowest overlap across all participants with only 4 out of 16 participants (25%). It is worth noting that this category included in-person contact which was only measured through the time that participants spent at self-identified social locations. From all comparisons (16 participants * 7 categories: N=112), a priori knowledge overlapped most often with what participants expected (65.2%; N=73), followed by an unexpected relation in the dataset (17.9%; N=20) and lastly an expected but not confirmed relation (17.0%; N=19).

V. DISCUSSION

Within the current study, first steps were taken towards the automatic detection and prediction of daily affective states in daily life. The study aimed at associating real-world variations across individuals in terms of behavior and physiology to individual variations in valence and arousal, using different types of models. Results showed that subject-dependent prediction models of valence and arousal outperformed on average leave-one-subject-out (LOSO) subject-independent

models. Spearman’s Rho correlations were on average 0.30 (valence) and 0.36 (arousal) for subject-dependent models in comparison to LOSO models with 0.17 (valence) and 0.28 (arousal). Results also indicated that individuals differ in terms of the number and relevance of features for arousal and valence predictions. The number of predictive features ranged from 1-9 for valence and 1-17 for arousal. The overlap between self-expected associations between valence and behavioral patterns was in most cases (N = 13; 81.25%) at least 4 out of 7 categories. When looking at the total number of comparisons, there was an overlap with a priori knowledge in 65.2% of all comparisons. These results give a first indication that exploiting such knowledge of individuals might help in the selection of to-be-collected data.

A. Discussion of main findings

Starting with the hypothesis that ‘subject-dependent models will outperform subject-independent (leave-one-subject-out) models in predicting affective states (valence and arousal) of individuals using passively sensed data’, predictions of valence and arousal levels were significantly higher for subject-dependent models. This result aligns with previous research focusing on short-term and intra-day levels, indicating that subject-dependent models outperform LOSO subject-independent models [72] and that person-specific stress detection models achieve a higher accuracy than general models [15], [16], [31], [32]. One reason for these results is that they take into account the variations and differences across humans (see [17]). The better performance of subject-dependent models over LOSO models also supports the idea of a tailored approach towards predictions instead of a “one-size-fits-all” approach [17]. In the study of [35], subject-independent prediction models of stress achieved a Spearman’s rho correlation of 0.19, which increased to 0.25 when accounting for individual differences in anxiety. In the current study, subject-dependent prediction models of both valence and arousal tend to outperform on average the model of [35], with 0.30 for valence and 0.36 for arousal. Additionally, the LOSO model of valence performed slightly worse (0.17) than [35] while the LOSO model of arousal performed slightly better (0.28) than [35]. The highest subject-dependent model of valence achieved a Spearman’s rho correlation of 0.60. For arousal, the highest subject-dependent Spearman correlation reached 0.69.

In terms of the second and third research question ‘What are the most important features for subject-dependent models of valence/arousal levels, and to what extent do these features differ across subject-dependent models?’, the results showed that individuals differ in the number and relevance of predictive features. These findings align with [22], indicating that the performance of subject-dependent prediction models differ across subjects and that using the same set of features and sensors for each individual may not be ideal. Individuals differed also both in terms of mean values and ranges of their self-reported affective state and objective measurements. This

TAFFC-2024-04-0220

is in line with previous findings indicating considerable differences in self-reports, biosignals across individuals, and smartphone usage (see [17]). Additionally, in the intake interview, it was reported by several participants that they sometimes evaluated their arousal according to their physical activity. This close connection between mental and physical arousal was also shown before (e.g. [98]). These results may also indicate that self-reports lack some strength as ground truth for prediction models, with one individual perceiving or interpreting arousal in a different way than another individual.

Regarding the fourth research question ‘To what extent does a priori self-evaluations of the relation between behavioral patterns and not feeling mentally correspond with subject-dependent models of valence?’, more than half of the categories mentioned in the intake interview were overlapping with the important features found in the dataset. However, there were also several cases where participants did not expect a change in their behavioral pattern when mentally not doing well while the data indicated a change. The mixed results of the current study might be due to participants having a lack of insight into their own behavioral patterns. Another reason might be a poor operationalization of or existing noise in these self-reported behavioral patterns.

B. Limitations of current work

When comparing the intake interview with the dataset, it became apparent that there was little overlap in the category of social contact and self-expected associations with not doing mentally well. One reason might be that most of the current features concern online behavior and interactions, thus not representing in-person social contacts very well with only one feature for this category (social location). Furthermore, since the COVID-19 pandemic, a lot of employees work partly from home and partly at their work locations, thus the capturing of GPS locations might not be fully reliable for the work-home distinction.

Due to technical difficulties with the smartphones and updates of Android, the AWARE app was not working equally well for all participants. In total, there were three participants (8, 10, 15) who experienced crashes of the app with valid data points equal or lower to 65%. An additional five participants (1, 4, 5, 7, 13) experienced irregular crashes of the AWARE app which resulted valid data points between 65% and 83%. This inconsistency in terms of AWARE data was mainly missing at random.

There were also connection problems with the Oura ring for two participants (5, 8) due to the type of smartphone provider. For another two participants (9, 11) the Oura ring stopped working and had to be replaced. For the former two participants, this may also have had an effect on their subject-dependent prediction performance. For the latter two participants, the data collection was continued for a longer time to balance the missing data. One limitation of this study is the scarcity of daytime cardiovascular measurements, which could serve as indicators for stress and potentially improve the

predictions of affective states. At present, it remains challenging to accurately interpret daytime stress measures due to the influence of individual activities and contextual factors (see e.g., [48]).

It was shown that participants differ in their reports of both valence and arousal, with a general tendency towards more pleasant and mentally active states (upper-right quadrant of valence-arousal grid). Since the current study focused on subject-dependent prediction models, a baselining approach was chosen due to the focus on relative changes within the individual. In this healthy population of participants, we assumed that large overall differences in scores between participants were caused mostly by differences in using the valence and arousal scales. We assumed that a right upward shift of a score was equally meaningful for a participant who reported to be almost always in the upper right quadrant, than for a participant whose scores were centered around the center of the valence-arousal grid. Note that this assumption only affected the subject-independent models as these predict the values of the left out subject, and not the subject-dependent models. By keeping the outcome measures the same across subject-dependent and subject-independent models, their performances could be compared. Additionally, the impact of other variables or factors such as personality [108] were not explored but may also have an impact on the way individuals fill in the daily questions about valence and arousal.

While the current study uses RF prediction models with continuous measure of valence and arousal, this does not necessarily mean that it is the best approach or classifier for all individuals. One reason why some of the subject-dependent prediction models of valence and arousal might have performed worse may be due to the nature of measurement. For some participants, behavioral patterns may be connected to more continuous measures but for other participants they may be connected more to categorical measures. Not each individual may report fine nuances of valence and arousal but rather reports on a day in a categorical manner (e.g., pleasant vs. neutral vs. unpleasant).

C. Implications for future work

The prediction of affective states in daily life can be complex and dependent on an individual, thus requiring a high multi-modal approach physiological data and contextual factors [42], [50]. Future research should collect data on predictive and outcome features at both intra- and inter-day levels to develop a more comprehensive understanding of the factors influencing affective states on micro and macro levels. In addition to increasing the frequency of daytime stress measurements, future research should explore ways to improve the interpretations and reliability of those measures. Additionally, it would be advisable to make use of a more robust smartphone application, preferably one that is usable by both Android and Apple users. Apps such as Beibe [109] or Insights [110] may provide a good alternative to this. The number and types of features should be extended, for instance, low privacy-invasive

TAFFC-2024-04-0220

features such as time of sunrise and weather were shown to have an impact on those predictions as well [35]. Regarding the category of social contact, it might be interesting to add more unobtrusive measures indicative of in-person contact, such as Bluetooth proximities (see [35]) or the number of social events mentioned in an individual's agenda. At the same time, the number of measured features should not be needlessly extended but selected according to the individual's background to avoid any harm to the individual's privacy.

Additionally, the Long Short-Term Memory (LSTM) approach has shown some promising results for continuous predictions of valence and arousal (see [22]). [22] has shown that LSTM models tend to outperform RF and AdaBoost, but that there are also cases where the other two work better. Thus, tailoring machine learning techniques to an individual is an advisable approach for future work as well. This includes comparing the prediction performance of categorical and continuous measures on a subject-dependent level to explore what works best for which individual. Another possibility would be to make use of Multitask Learning (MTL), as proposed by [67], [108], [111] highlights that individual differences can be accounted for with MTL when predicting mood, health and stress based on subjective (surveys) and objective measurements (wearable sensors, smartphone logs, weather).

One way to avoid a large burden on the participants but capturing important moments throughout or between days could be done with just-in-time EMA questions [112]. This would require a certain amount of training for each individual in order to identify deviations that can be used to trigger short questions about those deviations. Through this, one might also get a better grasp on durations and transitions of affective states. In practice, this would imply developing a monitoring system that learns about patterns from self-reports of the user in the initial months to create a subject-dependent prediction model. Subsequently, the monitoring system should be in a state to run unobtrusively and continuously in the background, and be sensitive enough to signal an individual when she or he might be at risk of developing prospective mental health problems. At that moment, EMA questions should be triggered to collect more detailed information about someone's mental state. With this more specific knowledge, personalized advice to prevent the development of psychological problems could be given. Attempts at detecting changes within-individuals were already made [42], [50], however, the results were limited to binary or binned stress predictions rather than continuous measures. The challenges of this approach are finding a suitable moment and manner to prompt an individual with questions or feedback [113] and avoid inducing additional stress with the feedback [114], [115]. Additionally, it will also have to be explored how subject-dependent models can be implemented and used in real-life settings. Subject-dependent models will require more data from individuals. Different methods and approaches of using a priori knowledge in the selection of features should be explored more in-depth. If successful, it could improve the efficiency of

models and lead to a set of relevant, high-quality features and data points.

For the current study, RF were used to investigate the predictive value of different objective and continuous features on subject-dependent and subject-independent levels across days over nine months. In future work, longer term dependencies between features and outcome variables are of particular interest. The study by [22] did make initial steps with continuous measures by testing time-series predictions with LTSM models since they are capable of learning long-term dependencies in the data. For example, the current features of physical activity can be combined and it can be compared how changes in physical activity are connected to short-term (e.g. daily) and long-term (e.g. weeks, months) changes in valence and arousal. With these long-term dependencies, it might thus become possible to move from individual predictive features (e.g. sedentary time, MVPA) connected to daily affective states to more stable behavioral patterns (e.g. physical activity) connected to affective states. For instance, noticeable changes in those stable patterns may be indicative of larger deviations in affective states. Consequently, just-in-time adaptive interventions (JITAI) could be connected and implemented according to those behavioral patterns and interdependencies [3].

There is still a gap between the methodology used in Social Sciences and Data Sciences, such as statistical analyses versus more advanced machine-learning techniques, both interpretation- and implementation-wise, future work should investigate whether these two can be combined and complement each other. For example, correlations are of particular interest for social science to understand the relationship between two or more variables. However, as stated by [116], predictor variables seen as important by machine learning techniques in terms of reducing classification error are not necessarily highly correlated with the outcome variables as indicated by statistical (linear) correlations.

Lastly, the current dataset provides an extensive amount of information on each participant through longitudinal measurements of affective states with EMA questions as well as behavioral and physiological patterns through various sensing modalities. Similar to [84], this dataset makes an addition to the sparse set of open, longitudinal datasets featuring subjective, continuous and objective measurement.

VI. CONCLUSION

Grasping the dynamics of affective states and their prediction in the wild using passive sensing remains challenging, however, with the continuous improvement of multimodal sensors it may eventually be possible to detect changes in mental health and intervene on time. The current study showed that predictions of affective states in daily life with continuous physiological and behavioral data from sensors can be improved with subject-dependent prediction models. However, the extent of improvement strongly depends on the individual. Subject-dependent prediction models of fine-grained valence levels

TAFFC-2024-04-0220

achieved on average a Spearman's rho of 0.28 [0.14 - 0.60] and prediction models of fine-grained arousal 0.36 [0.16 - 0.69]. Variations across subjects were shown in terms of affect perception, feature importance and model performance. Subject-dependent prediction models of valence and arousal outperformed leave-one-subject-out subject-independent prediction models. First results of comparing subjective with objective measures of behavioral patterns related to valence show that, for many cases a priori knowledge might help to reduce the number and type of collected data per person. The presented results advance the field of affective state prediction by focusing and highlighting the importance of individual differences and their impact on prediction models of valence and arousal. Future work should continue to focus on subject-dependent prediction models and explore more extensively how prediction models can be tailored to the individual to improve model performance. These models could then be used as input for the development and training of unobtrusive and continuous subject-dependent monitoring systems of affective states.

DATA AVAILABILITY STATEMENT

The dataset that supports the findings will be pseudo-anonymized and made publicly available in a separate publication. The authors are currently preparing a separate journal article on this matter and expect the article to be published in 2025.

ACKNOWLEDGMENT

The authors would like to thank all participants who delivered extensive and important data for this paper and the associated analyses.

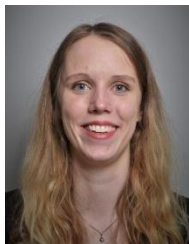
REFERENCES

- [1] P. P. Morita, "Chapter 5 - Design of mobile health technology," in *Design for Health*, A. Sethumadhavan and F. Sasangohar Eds.: Academic Press, 2020, pp. 87-102.
- [2] L. Koinis, R. J. Mobbs, R. D. Fonseka, and P. Natarajan, "A commentary on the potential of smartphones and other wearable devices to be used in the identification and monitoring of mental illness," (in eng), *Ann Transl Med*, vol. 10, no. 24, p. 1420, Dec 2022, doi: 10.21037/atm-21-6016.
- [3] L. Wang and L. C. Miller, "Just-in-the-Moment Adaptive Interventions (JITAI): A Meta-Analytical Review," *Health Communication*, vol. 35, no. 12, pp. 1531-1544, 2020/10/14 2020, doi: 10.1080/10410236.2019.1652388.
- [4] J. Zhou, B. Lamichhane, D. Ben-Zeev, A. Campbell, and A. Sano, "Predicting Psychotic Relapse in Schizophrenia With Mobile Sensor Data: Routine Cluster Analysis," *JMIR Mhealth Uhealth*, vol. 10, no. 4, p. e31006, 2022/4/11 2022, doi: 10.2196/31006.
- [5] H. F. v. d. Molen, K. Nieuwenhuijsen, M. H. W. Frings-Dresen, and G. d. Groene, "Work-related psychosocial risk factors for stress-related mental disorders: an updated systematic review and meta-analysis," *BMJ Open*, vol. 10, no. 7, p. e034849, 2020, doi: 10.1136/bmjopen-2019-034849.
- [6] F. Chirico, "Job stress models for predicting burnout syndrome: a review," (in eng), *Ann Ist Super Sanita*, vol. 52, no. 3, pp. 443-456, Jul-Sep 2016, doi: 10.4415/ann_16_03_17.
- [7] J. Li, M. Zhang, A. Loerbroks, P. Angerer, and J. Siegrist, "Work stress and the risk of recurrent coronary heart disease events: A systematic review and meta-analysis," *International Journal of Occupational Medicine and Environmental Health*, 2014/11/08 2014, doi: 10.2478/s13382-014-0303-7.
- [8] P. A. Collins and A. C. C. Gibbs, "Stress in police officers: a study of the origins, prevalence and severity of stress-related symptoms within a county police force," *Occupational Medicine*, vol. 53, no. 4, pp. 256-264, 2003, doi: 10.1093/ocmed/kqg061.
- [9] J. J. Gross, H. Uusberg, and A. Uusberg, "Mental illness and well-being: an affect regulation perspective," *World Psychiatry*, vol. 18, no. 2, pp. 130-139, 2019, doi: <https://doi.org/10.1002/wps.20618>.
- [10] J. J. Gross and H. Jazaieri, "Emotion, Emotion Regulation, and Psychopathology: An Affective Science Perspective," *Clinical Psychological Science*, vol. 2, no. 4, pp. 387-401, 2014, doi: 10.1177/2167702614536164.
- [11] J. J. Gross and R. F. Muñoz, "Emotion regulation and mental health," *Clinical Psychology: Science and Practice*, vol. 2, no. 2, pp. 151-164, 1995, doi: 10.1111/j.1468-2850.1995.tb00036.x.
- [12] T. J. Trull and U. Ebner-Priemer, "The Role of Ambulatory Assessment in Psychological Science," *Current Directions in Psychological Science*, vol. 23, no. 6, pp. 466-470, 2014, doi: 10.1177/0963721414550706.
- [13] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161-1178, 1980, doi: 10.1037/h0077714.
- [14] A. Durmaz, İ. Dursun, and E. T. Kabadayi, "Mitigating the Effects of Social Desirability Bias in Self-Report Surveys: Classical and New Techniques," in *Applied Social Science Approaches to Mixed Methods Research*, M. L. Baran and J. E. Jones Eds. Hershey, PA, USA: IGI Global, 2020, pp. 146-185.
- [15] G. Kjellsson, P. Clarke, and U.-G. Gerdtham, "Forgetting to remember or remembering to forget: A study of the recall period length in health care survey questions," *Journal of Health Economics*, vol. 35, pp. 34-46, 2014/05/01/ 2014, doi: <https://doi.org/10.1016/j.jhealeco.2014.01.007>.
- [16] M. E. Short *et al.*, "How accurate are self-reports? Analysis of self-reported health care utilization and absence when compared with administrative data," (in eng), *J Occup Environ Med*, vol. 51, no. 7, pp. 786-96, Jul 2009, doi: 10.1097/JOM.0b013e3181a86671.
- [17] E. Vildjiounaite *et al.*, "Challenges of learning human digital twin: case study of mental wellbeing: Using sensor data and machine learning to create HDT," presented at the Proceedings of the 16th International Conference on Pervasive Technologies Related to Assistive Environments, Corfu, Greece, 2023. [Online]. Available: <https://doi.org/10.1145/3594806.3596538>.
- [18] Y. L. Hsu, J. S. Wang, W. C. Chiang, and C. H. Hung, "Automatic ECG-Based Emotion Recognition in Music Listening," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 85-99, 2020, doi: 10.1109/TAFFC.2017.2781732.
- [19] L. Romeo, A. Cavallo, L. Pepa, N. Bianchi-Berthouze, and M. Pontil, "Multiple Instance Learning for Emotion Recognition Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 389-407, 2022, doi: 10.1109/TAFFC.2019.2954118.
- [20] F. Li, K. Shirahama, M. A. Nisar, X. Huang, and M. Grzegorzec, "Deep Transfer Learning for Time Series Data Based on Sensor Modality Classification," *Sensors*, vol. 20, no. 15, p. 4271, 2020. [Online]. Available: <https://www.mdpi.com/1424-8220/20/15/4271>.
- [21] B. Zhao, Z. Wang, Z. Yu, and B. Guo, "EmotionSense: Emotion Recognition Based on Wearable Wristband," in *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI)*, 8-12 Oct. 2018 2018, pp. 346-355, doi: 10.1109/SmartWorld.2018.00091.
- [22] P. Siirtola, S. Tamminen, G. Chandra, A. Ihalapathirana, and J. Röning, "Predicting Emotion with Biosignals: A Comparison of Classification and Regression Models for Estimating Valence and Arousal Level Using Wearable Sensors," *Sensors*, vol. 23, no. 3, p. 1598, 2023. [Online]. Available: <https://www.mdpi.com/1424-8220/23/3/1598>.
- [23] B. Helgadóttir, Y. Forsell, and Ö. Eklom, "Physical Activity Patterns of People Affected by Depressive and Anxiety Disorders as Measured by Accelerometers: A Cross-Sectional Study," *PLOS ONE*, vol. 10, no. 1, p. e0115894, 2015, doi: 10.1371/journal.pone.0115894.
- [24] M. Hagströmer, P. Oja, and M. Sjöström, "Physical activity and inactivity in an adult population assessed by accelerometry," (in eng), *Med Sci Sports Exerc*, vol. 39, no. 9, pp. 1502-1508, 2007/09// 2007, doi: 10.1249/mss.0b013e3180a76de5.
- [25] L. Canzian and M. Musolesi, "Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis," presented at the Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Osaka, Japan, 2015. [Online]. Available: <https://doi.org/10.1145/2750858.2805845>.
- [26] S. Saeb, E. G. Lattie, S. M. Schueller, K. P. Kording, and D. C. Mohr, "The relationship between mobile phone location sensor data and

- depressive symptom severity," (in eng), *PeerJ*, vol. 4, p. e2537, 2016, doi: 10.7717/peerj.2537.
- [27] E. Smets *et al.*, "Large-scale wearable data reveal digital phenotypes for daily-life stress detection," *npj Digital Medicine*, vol. 1, no. 1, p. 67, 2018/12/12 2018, doi: 10.1038/s41746-018-0074-9.
- [28] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *Journal of Biomedical Informatics*, vol. 73, pp. 159-170, 2017/09/01/ 2017, doi: <https://doi.org/10.1016/j.jbi.2017.08.006>.
- [29] P. Schmidt, A. Reiss, R. Dürichen, and K. V. Laerhoven, "Wearable-Based Affect Recognition—A Review," *Sensors*, vol. 19, no. 19, p. 4079, 2019. [Online]. Available: <https://www.mdpi.com/1424-8220/19/19/4079>.
- [30] A. Alberdi, A. Aztiria, and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review," *Journal of Biomedical Informatics*, vol. 59, pp. 49-75, 2016/02/01/ 2016, doi: <https://doi.org/10.1016/j.jbi.2015.11.007>.
- [31] E. Smets, W. D. Raedt, and C. V. Hoof, "Into the Wild: The Challenges of Physiological Stress Detection in Laboratory and Ambulatory Settings," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 463-473, 2019, doi: 10.1109/JBHI.2018.2883751.
- [32] E. Smets *et al.*, "Comparison of Machine Learning Techniques for Psychophysiological Stress Detection," in *Pervasive Computing Paradigms for Mental Health*, Cham, S. Serino, A. Matic, D. Giakoumis, G. Lopez, and P. Cipresso, Eds., 2016// 2016: Springer International Publishing, pp. 13-22.
- [33] C. da Estrela, S. MacNeil, and J.-P. Gouin, "Heart rate variability moderates the between- and within-person associations between daily stress and negative affect," *International Journal of Psychophysiology*, vol. 162, pp. 79-85, 2021/04/01/ 2021, doi: <https://doi.org/10.1016/j.ijpsycho.2021.02.001>.
- [34] J. Tervonen *et al.*, "Personalized mental stress detection with self-organizing map: From laboratory to the field," *Computers in Biology and Medicine*, vol. 124, p. 103935, 2020/09/01/ 2020, doi: <https://doi.org/10.1016/j.combiomed.2020.103935>.
- [35] B. M. Booth, H. Vrzakova, S. M. Mattingly, G. J. Martinez, L. Faust, and S. K. D'Mello, "Toward Robust Stress Prediction in the Age of Wearables: Modeling Perceived Stress in a Longitudinal Study With Information Workers," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2201-2217, 2022, doi: 10.1109/TAFFC.2022.3188006.
- [36] A. Muaremi, B. Arnrich, and G. Tröster, "Towards Measuring Stress with Smartphones and Wearable Devices During Workday and Sleep," *BioNanoScience*, vol. 3, no. 2, pp. 172-183, 2013/06/01 2013, doi: 10.1007/s12668-013-0089-2.
- [37] D. A. Adler *et al.*, "Burnout and the Quantified Workplace: Tensions around Personal Sensing Interventions for Stress in Resident Physicians," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. CSCW2, p. Article 430, 2022, doi: 10.1145/3555531.
- [38] D. C. Mohr, K. Shilton, and M. Hotopf, "Digital phenotyping, behavioral sensing, or personal sensing: names and transparency in the digital age," *npj Digital Medicine*, vol. 3, no. 1, p. 45, 2020/03/25 2020, doi: 10.1038/s41746-020-0251-5.
- [39] C. Ingraham, "An insurance company wants you to hand over your Fitbit data so it can make more money. Should you?," vol. 2023, ed: Washington Post, 2018.
- [40] H. Jazaieri, H. L. Urry, and J. J. Gross, "Affective Disturbance and Psychopathology: An Emotion Regulation Perspective," *Journal of Experimental Psychopathology*, vol. 4, no. 5, pp. 584-599, 2013, doi: 10.5127/jep.030312.
- [41] A. Zenonos, A. Khan, G. Kalogridis, S. Vatsikas, T. Lewis, and M. Sooriyabandara, "HealthyOffice: Mood recognition at work using smartphones and wearable sensors," in *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, 14-18 March 2016 2016, pp. 1-6, doi: 10.1109/PERCOMW.2016.7457166.
- [42] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous stress detection using a wrist device: in laboratory and real life," presented at the Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, Heidelberg, Germany, 2016. [Online]. Available: <https://doi.org/10.1145/2968219.2968306>.
- [43] P. Siirtola, "Continuous stress detection using the sensors of commercial smartwatch," presented at the Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, London, United Kingdom, 2019. [Online]. Available: <https://doi.org/10.1145/3341162.3344831>.
- [44] Y. S. Can, B. Arnrich, and C. Ersoy, "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey," *Journal of Biomedical Informatics*, vol. 92, p. 103139, 2019/04/01/ 2019, doi: <https://doi.org/10.1016/j.jbi.2019.103139>.
- [45] S. Gedam and S. Paul, "Automatic Stress Detection Using Wearable Sensors and Machine Learning: A Review," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1-3 July 2020 2020, pp. 1-7, doi: 10.1109/ICCCNT49239.2020.9225692.
- [46] A. Shaw, N. Simsiri, I. Deznaby, M. Fiterau, and T. Rahaman, "Personalized Student Stress Prediction with Deep Multitask Network," p. arXiv:1906.11356doi: 10.48550/arXiv.1906.11356.
- [47] H. de Vries, H. Oldenhuis, C. van der Schans, R. Sanderman, and W. Kamphuis, "Does Wearable-Measured Heart Rate Variability During Sleep Predict Perceived Morning Mental and Physical Fitness?," *Applied Psychophysiology and Biofeedback*, vol. 48, no. 2, pp. 247-257, 2023/06/01 2023, doi: 10.1007/s10484-022-09578-8.
- [48] H. J. de Vries, H. J. M. Pennings, C. P. van der Schans, R. Sanderman, H. K. E. Oldenhuis, and W. Kamphuis, "Wearable-Measured Sleep and Resting Heart Rate Variability as an Outcome of and Predictor for Subjective Stress Measures: A Multiple N-of-1 Observational Study," *Sensors*, vol. 23, no. 1, p. 332, 2023. [Online]. Available: <https://www.mdpi.com/1424-8220/23/1/332>.
- [49] H. de Vries, W. Kamphuis, C. van der Schans, R. Sanderman, and H. Oldenhuis, "Trends in Daily Heart Rate Variability Fluctuations Are Associated with Longitudinal Changes in Stress and Somatisation in Police Officers," *Healthcare*, vol. 10, no. 1, p. 144, 2022. [Online]. Available: <https://www.mdpi.com/2227-9032/10/1/144>.
- [50] V. Mishra *et al.*, "Investigating the Role of Context in Perceived Stress Detection in the Wild," presented at the Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, Singapore, Singapore, 2018. [Online]. Available: <https://doi.org/10.1145/3267305.3267537>.
- [51] Y. S. Can, D. Gokay, D. R. Kılıç, D. Ekiz, N. Chalabianloo, and C. Ersoy, "How Laboratory Experiments Can Be Exploited for Monitoring Stress in the Wild: A Bridge Between Laboratory and Daily Life," *Sensors*, vol. 20, no. 3, p. 838, 2020. [Online]. Available: <https://www.mdpi.com/1424-8220/20/3/838>.
- [52] V. Engert, S. I. Efanov, A. Duchesne, S. Vogel, V. Corbo, and J. C. Pruessner, "Differentiating anticipatory from reactive cortisol responses to psychosocial stress," *Psychoneuroendocrinology*, vol. 38, no. 8, pp. 1328-1337, 2013/08/01/ 2013, doi: <https://doi.org/10.1016/j.psyneuen.2012.11.018>.
- [53] P. E. Velmovitsky, P. Alencar, S. T. Leatherdale, D. Cowan, and P. P. Morita, "Using apple watch ECG data for heart rate variability monitoring and stress prediction: A pilot study," (in English), *Frontiers in Digital Health*, Original Research vol. 4, 2022-December-09 2022, doi: 10.3389/fdgh.2022.1058826.
- [54] P. E. Velmovitsky, "Use of Smart Technology Tools for Supporting Public Health Surveillance: From Development of a Mobile Health Platform to Application in Stress Prediction," 2023.
- [55] M. Teychenne, L. D. Stephens, S. A. Costigan, D. L. Olstad, B. Stubbs, and A. I. Turner, "The association between sedentary behaviour and indicators of stress: a systematic review," *BMC Public Health*, vol. 19, no. 1, p. 1357, 2019/10/23 2019, doi: 10.1186/s12889-019-7717-x.
- [56] D. Bhugra and A. Mastrogiani, "Globalisation and mental disorders: Overview with relation to depression," *The British Journal of Psychiatry*, vol. 184, no. 1, pp. 10-20, 2004, doi: 10.1192/bjp.184.1.10.
- [57] K. N. Anderson and A. J. Bradley, "Sleep disturbance in mental health problems and neurodegenerative disease," *Nature and Science of Sleep*, vol. 5, no. null, pp. 61-75, 2013/05/31 2013, doi: 10.2147/NSS.S34842.
- [58] D. Currier, R. Lindner, M. J. Spittal, S. Cvetkovski, J. Pirakis, and D. R. English, "Physical activity and depression in men: Increased activity duration and intensity associated with lower likelihood of current depression," *Journal of Affective Disorders*, vol. 260, pp. 426-431, 2020/01/01/ 2020, doi: <https://doi.org/10.1016/j.jad.2019.09.061>.
- [59] M. Teychenne, K. Ball, and J. Salmon, "Associations between physical activity and depressive symptoms in women," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 5, no. 1, p. 27, 2008/05/06 2008, doi: 10.1186/1479-5868-5-27.

- [60] M. T. De Mello, V. d. A. Lemos, H. K. M. Antunes, L. Bittencourt, R. Santos-Silva, and S. Tufik, "Relationship between physical activity and depression and anxiety symptoms: A population study," *Journal of Affective Disorders*, vol. 149, no. 1, pp. 241-246, 2013/07/01/ 2013, doi: <https://doi.org/10.1016/j.jad.2013.01.035>.
- [61] S. Saeb *et al.*, "Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study," *J Med Internet Res*, vol. 17, no. 7, p. e175, 2015/07/15 2015, doi: 10.2196/jmir.4273.
- [62] F. Wahle, T. Kowatsch, E. Fleisch, M. Rufer, and S. Weidt, "Mobile Sensing and Support for People With Depression: A Pilot Trial in the Wild," *JMIR Mhealth Uhealth*, vol. 4, no. 3, p. e111, 2016/09/21 2016, doi: 10.2196/mhealth.5960.
- [63] K. Opoku Asare *et al.*, "Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status: A longitudinal data analysis," *Pervasive and Mobile Computing*, vol. 83, p. 101621, 2022/07/01/ 2022, doi: <https://doi.org/10.1016/j.pmcj.2022.101621>.
- [64] I. Moshe *et al.*, "Predicting Symptoms of Depression and Anxiety Using Smartphone and Wearable Data," (in English), *Frontiers in Psychiatry, Original Research* vol. 12, 2021-January-28 2021, doi: 10.3389/fpsy.2021.625247.
- [65] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. V. Laerhoven, "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection," presented at the Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 2018. [Online]. Available: <https://doi.org/10.1145/3242969.3242985>.
- [66] S. G. Hofmann, "Toward a Cognitive-Behavioral Classification System for Mental Disorders," *Behavior Therapy*, vol. 45, no. 4, pp. 576-587, 2014/07/01/ 2014, doi: <https://doi.org/10.1016/j.beth.2014.03.001>.
- [67] S. Taylor, N. Jacques, E. Nosakhare, A. Sano and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health." *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 200-213, 2020, <https://doi.org/10.1109/taffc.2017.2784832>.
- [68] A. Ruf, A. B. Neubauer, E. D. Koch, U. Ebner-Priemer, A. Reif, and S. Matura, "Individual differences in the dietary response to stress in ecological momentary assessment: Does the individual-difference model need expansion?," *Applied Psychology: Health and Well-Being*, vol. 15, no. 2, pp. 629-649, 2023, doi: <https://doi.org/10.1111/aphw.12400>.
- [69] J. L. Hicks *et al.*, "Best practices for analyzing large-scale health data from wearables and smartphone apps," *npj Digital Medicine*, vol. 2, no. 1, p. 45, 2019/06/03 2019, doi: 10.1038/s41746-019-0121-1.
- [70] P. E. Velmovitsky, T. Bevilacqua, P. Alencar, D. Cowan, and P. P. Morita, "Convergence of Precision Medicine and Public Health Into Precision Public Health: Toward a Big Data Perspective," (in English), *Frontiers in Public Health, Review* vol. 9, 2021-April-06 2021, doi: 10.3389/fpubh.2021.561873.
- [71] L. Berkemeier *et al.*, "Towards continuous mental state detection in everyday settings: investigating between-subjects variations in a longitudinal study," in *2023 Applied Human Factors and Ergonomics (AHFE) International: Neuroergonomics and Cognitive Engineering*, 20-24 July 2023, pp. 131-140, doi: 10.54941/ahfe1003013.
- [72] M. T. Uddin, G. Zamzmi and S. J. Canavan, "Cooperative learning for personalized context-aware pain assessment from Wearable Data." *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 11, pp. 5260-5271, 2023, <https://doi.org/10.1109/jbhi.2023.3294903>.
- [73] K. Levecque, F. Anseel, A. De Beuckelaer, J. Van der Heyden, and L. Gisle, "Work organization and mental health problems in PhD students," *Research Policy*, vol. 46, no. 4, pp. 868-879, 2017/05/01/ 2017, doi: <https://doi.org/10.1016/j.respol.2017.02.008>.
- [74] D. Ferreira, V. Kostakos, and A. K. Dey, "AWARE: Mobile Context Instrumentation Framework," (in English), *Frontiers in ICT, Technology Report* vol. 2, 2015-April-20 2015, doi: 10.3389/fict.2015.00006.
- [75] R. Cao *et al.*, "Accuracy Assessment of Oura Ring Nocturnal Heart Rate and Heart Rate Variability in Comparison With Electrocardiography in Time and Frequency Domains: Comprehensive Analysis," *J Med Internet Res*, vol. 24, no. 1, p. e27487, 2022/1/18 2022, doi: 10.2196/27487.
- [76] M. Altini, and H. Kinnunen, "The promise of sleep: A multi-sensor approach for accurate sleep stage detection using the Oura Ring." *Sensors*, vol. 21, no. 13, p. 4302, 2021, <https://doi.org/10.3390/s21134302>.
- [77] M. de Zambotti, N. Celini, L. Menghini, M. Sarlo and F. C. Baker, "The sleep of the ring: Comparison of the oura sleep tracker against polysomnography." *Behavioral Sleep Medicine*, vol. 17, no. 2, pp. 124-136, 2017, <https://doi.org/10.1080/15402002.2017.1300587>.
- [78] T. Svensson, K. Madhawa, H. NT, U. Chung and A. K. Svensson, "Validity and reliability of the Oura Ring Generation 3 (gen3) with Oura Sleep Staging Algorithm 2.0 (OSSA 2.0) when compared to multi-night ambulatory polysomnography: A validation study of 96 participants and 421,045 epochs." *Sleep Medicine*, vol. 115, , pp. 251-263, 2024 <https://doi.org/10.1016/j.sleep.2024.01.020>.
- [79] E. Kristiansson, J. Fridolfsson, D. Arvidsson, A. Holmäng, M. Börjesson and U. Andersson-Hall, "Validation of oura ring energy expenditure and steps in laboratory and free-living." *BMC Medical Research Methodology*, vol. 23, no. 1, 2023, <https://doi.org/10.1186/s12874-023-01868-x>.
- [80] S. Jafarlou *et al.* "Objective prediction of next-day's affect using multimodal physiological and behavioral data: Algorithm development and Validation Study." *JMIR Formative Research*, vol. 7, 2023 <https://doi.org/10.2196/39425>.
- [81] M. Price *et al.* "A large clinical trial to improve well-being during the transition to college using wearables: The lived experiences measured using rings study." *Contemporary Clinical Trials*, vol. 133, p. 107338, 2023 <https://doi.org/10.1016/j.cct.2023.107338>.
- [82] TNO. "How am I onderzoek, design-kit." <https://www.monitorarbeid.tno.nl/nl-nl/coronacrisis/howami-onderzoek/> (accessed 09-10, 2023).
- [83] L. Braunsman, F. Beermann, H. K. Strüder and V. Abeln, "Self-selected versus imposed running intensity and the acute effects on mood, cognition, and (a)periodic brain activity." *Cognitive Neurodynamics*, 2024, <https://doi.org/10.1007/s11571-024-10084-2>.
- [84] S. Kang *et al.* "K-EmoPhone: A Mobile and wearable dataset with in-situ emotion, stress, and attention labels." *Scientific Data*, vol. 10, no. 1, 2023, <https://doi.org/10.1038/s41597-023-02248-2>.
- [85] H. G. Kim, E. J. Cheon, D. S. Bai, Y. H. Lee, and B. H. Koo, "Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature," (in eng), *Psychiatry Investig*, vol. 15, no. 3, pp. 235-245, Mar 2018, doi: 10.30773/pi.2017.08.17.
- [86] C. Schubert, M. Lambertz, R. A. Nelesen, W. Bardwell, J. B. Choi, and J. E. Dimsdale, "Effects of stress on heart rate complexity—A comparison between short-term and chronic stress." *Biological Psychology*, vol. 80, no. 3, pp. 325-332, 2009/03/01/ 2009, doi: <https://doi.org/10.1016/j.biopsycho.2008.11.005>.
- [87] E. M. Jackson, "STRESS RELIEF: The Role of Exercise in Stress Management," *ACSM's Health & Fitness Journal*, vol. 17, no. 3, 2013. [Online]. Available: https://journals.lww.com/acsm-healthfitness/fulltext/2013/05000/stress_relief_the_role_of_exercise_in_stress.6.aspx.
- [88] R. Wang *et al.*, "StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones," presented at the Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Seattle, Washington, 2014. [Online]. Available: <https://doi.org/10.1145/2632048.2632054>.
- [89] L. E. Charles *et al.*, "Association of perceived stress with sleep duration and sleep quality in police officers," (in eng), *Int J Emerg Ment Health*, vol. 13, no. 4, pp. 229-41, 2011.
- [90] C. R. Rebello, P. B. Kallingappa, and P. G. Hegde, "Assessment of perceived stress and association with sleep quality and attributed stressors among 1(st)-year medical students: A cross-sectional study from Karwar, Karnataka, India," (in eng), *Ci Ji Yi Xue Za Zhi*, vol. 30, no. 4, pp. 221-226, Oct-Dec 2018, doi: 10.4103/tcmj.tcmj_85_17.
- [91] S. Yoon, S.-s. Lee, J.-m. Lee, and K. Lee, "Understanding notification stress of smartphone messenger app," presented at the CHI '14 Extended Abstracts on Human Factors in Computing Systems, Toronto, Ontario, Canada, 2014. [Online]. Available: <https://doi.org/10.1145/2559206.2581167>.
- [92] V. Apaolaza, P. Hartmann, C. D'Souza, and A. Gilsanz, "Mindfulness, Compulsive Mobile Social Media Use, and Derived Stress: The Mediating Roles of Self-Esteem and Social Anxiety," *Cyberpsychology, Behavior, and Social Networking*, vol. 22, no. 6, pp. 388-396, 2019/06/01 2019, doi: 10.1089/cyber.2018.0681.
- [93] M. Pielot and L. Rello, "Productive, anxious, lonely: 24 hours without push notifications," presented at the Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, Vienna, Austria, 2017. [Online]. Available: <https://doi.org/10.1145/3098279.3098526>.
- [94] R. Wang *et al.*, "Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing," *Proc. ACM Interact. Mob.*

- Wearable Ubiquitous Technol.*, vol. 2, no. 1, p. Article 43, 2018, doi: 10.1145/3191775.
- [95] S. Saeb, Z. Mi, M. Kwasny, C. J. Karr, K. Kording, and D. C. Mohr, "The relationship between clinical, momentary, and sensor-based assessment of depression," in *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, 20-23 May 2015 2015, pp. 229-232, doi: 10.4108/icst.pervasivehealth.2015.259034.
- [96] S. R. Müller, H. Peters, S. C. Matz, W. Wang, and G. M. Harari, "Investigating the Relationships between Mobility Behaviours and Indicators of Subjective Well-Being Using Smartphone-Based Experience Sampling and GPS Tracking," *European Journal of Personality*, vol. 34, no. 5, pp. 714-732, 2020, doi: 10.1002/per.2262.
- [97] K. Wong, A. H. S. Chan, and S. C. Ngan, "The Effect of Long Working Hours and Overtime on Occupational Health: A Meta-Analysis of Evidence from 1998 to 2018," *International Journal of Environmental Research and Public Health*, vol. 16, no. 12, p. 2102, 2019. [Online]. Available: <https://www.mdpi.com/1660-4601/16/12/2102>.
- [98] K. L. Zellars, J. A. Meurs, P. L. Perrewé, C. J. Kacmar, and A. M. Rossi, "Reacting to and recovering from a stressful situation: The negative affectivity-physiological arousal relationship," *Journal of Occupational Health Psychology*, vol. 14, no. 1, pp. 11-22, 2009, doi: 10.1037/a0013823.
- [99] J.J., Allaire, J.J. "RStudio: Integrated Development for R.; RStudio", <http://www.rstudio.com>. (accessed 04-04. 2024).
- [100] L. Breiman, "Random Forests", *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2021, <https://doi.org/10.1023/a:1010933404324>.
- [101] L. Grinsztajn, E. Oyallon and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?", in *2022 Conference on Neural Information Processing Systems: Track on Datasets and Benchmarks*, doi:10.48550/arXiv.2207.08815.
- [102] P. Probst, M. N. Wright and A.-L. Boulesteix, "Hyperparameters and tuning strategies for Random Forest." *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 3, 2019, <https://doi.org/10.1002/widm.1301>.
- [103] E. Scornet, "Tuning parameters in random forests." *ESAIM: Proceedings and Surveys*, vol. 60, pp. 144-162, 2017, <https://doi.org/10.1051/proc/201760144>.
- [104] L. A. Jeni, J. F. Cohn and F. de la Torre, "Facing imbalanced data-recommendations for the use of performance metrics.", in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 245-251, <https://doi.org/10.1109/acii.2013.47>.
- [105] A. Jovic, K. Brkic and N. Bogunovic, "A review of feature selection methods with applications.", in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1200-1205, <https://doi.org/10.1109/mipro.2015.7160458>.
- [106] P. Glaysheer, J. M. Katzy and S. An., "Iterative Subtraction Method for Feature Ranking." *arXiv.Org*, 2019, doi: arxiv.org/abs/1906.05718v1.
- [107] S. O'Hara et al. "Iterative feature removal yields highly discriminative pathways.", *BMC Genomics*, vol. 14, no. 1, 2013, <https://doi.org/10.1186/1471-2164-14-832>.
- [108] N. Jacques, O. Rudovic, S. Taylor, A. Sano and R. Picard, "Predicting Tomorrow's Mood, Health, and Stress Level using Personalized Multitask Learning and Domain Adaptation", in *2017 Proceedings of Machine Learning Research*, pp. 17-33.
- [109] J. Torous, M. V. Kiang, J. Lorme, and J.-P. Onnela, "New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research," *JMIR Mental Health*, vol. 3, no. 2, p. e16, 2016/05/05 2016, doi: 10.2196/mental.5165.
- [110] C. Montag, H. Baumeister, C. Kannen, R. Sariyska, E.-M. Meßner, and M. Brand, "Concept, Possibilities and Pilot-Testing of a New Smartphone Application for the Social and Life Sciences to Study Human Behavior Including Validation Data from Personality Psychology," *J*, vol. 2, no. 2, pp. 102-115, 2019. [Online]. Available: <https://www.mdpi.com/2571-8800/2/2/8>.
- [111] H. Yu and A. Sano, "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)*, pp. 5884-5887, <https://doi.org/10.1109/embc44109.2020.9176242>.
- [112] S. Schneider, D. U. Jungheanel, J. M. Smyth, C. K. Fred Wen, and A. A. Stone, "Just-in-time adaptive ecological momentary assessment (JITA-EMA)," *Behavior Research Methods*, 2023/02/25 2023, doi: 10.3758/s13428-023-02083-8.
- [113] C. M. Kelley and A. C. McLaughlin, "Individual Differences in the Benefits of Feedback for Learning," *Human Factors*, vol. 54, no. 1, pp. 26-35, 2012, doi: 10.1177/0018720811423919.
- [114] B. Yu, M. Funk, J. Hu, Q. Wang, and L. Feijs, "Biofeedback for Everyday Stress Management: A Systematic Review," (in English), *Frontiers in ICT*, Review vol. 5, 2018-September-07 2018, doi: 10.3389/fict.2018.00023.
- [115] N. Moraveji, A. Adishesan, and T. Hagiwara, "BreathTray: augmenting respiration self-regulation without cognitive deficit," presented at the CHI '12 Extended Abstracts on Human Factors in Computing Systems, Austin, Texas, USA, 2012. [Online]. Available: <https://doi.org/10.1145/2212776.2223810>.
- [116] A. Bauranov, S. Parks, X. Jiang, J. Rakas and M. C. González, "Quantifying the resilience of the U.S. Domestic Aviation Network during the COVID-19 pandemic." *Frontiers in Built Environment*, vol. 7, 2021, <https://doi.org/10.3389/fbuil.2021.642295>.



Lea Berkemeier is a PhD candidate in the Department of Psychology, Health and Technology at the University of Twente. She has a background in Human Factors and Engineering Psychology (MSc) and Personalized Monitoring and Coaching (MSc). Her research focuses on the automatic monitoring and prediction of mental health and lies at the intersection of

Data Science, Health Sciences and Psychology. By combining subjective measurements with data from passive sensors such as wearables and smartphones, she explores how (changes in) affective states can be monitored in real life, to what extent these processes differ across individuals and how machine learning can be used for predictions of mental health on the long run.



Wim Kamphuis received his PhD in Social and Organizational Psychology at Tilburg University, the Netherlands. Currently, he is a Senior Scientist at TNO, the Netherlands Organization for Applied Scientific Research, in the department of Learning and Workforce Development. His research revolves around the themes of resilience, stress, and coping of military

and law enforcement personnel. At present, he leads programs and projects focused on the application of wearable and app technology for continuous (automatic) monitoring of mental and physical states, and the development of models and predictive algorithms in this field, resulting in personal monitoring systems tailored for high-risk professionals.



Anne-Marie Brouwer is a senior scientist at TNO Netherlands Applied Research Organization, with a joint appointment as a full professor 'Mental State Monitoring' at the Radboud University Nijmegen. Her research focuses on extracting information about cognitive and emotional states from implicit, continuous signals such as EEG brain signals and eye movement behavior,

and how this information can be sensibly and responsibly harnessed in applications that benefit people's performance and well-being in different domains. Prof. Brouwer published over

TAFFC-2024-04-0220

150 journal and proceeding papers, and is chief editor of the Consumer Neuroergonomics section of Frontiers in Neuroergonomics.



Herman J. de Vries has a background in physiotherapy and clinical health sciences and received his PhD on modeling mental resilience using wearables and apps at the University of Groningen. His current efforts as a research scientist at TNO, the Netherlands Organization for Applied Scientific Research, focus on optimizing the health and well-being of individuals in

high-risk professions (e.g., military and police) using digital health tools such as wearable sensors and self-monitoring technologies.



Maarten Schadd holds a MSc and a PhD degree in Artificial Intelligence and is an expert in Reinforcement Learning, Monte-Carlo techniques and optimization algorithms. His main research focus at TNO, the Netherlands Organization for Applied Scientific Research, is on Artificial Intelligence for decision support

systems in the military domain. He has gained experience in developing complex interacting systems in the private sector.



Jan Ubbo van Baardewijk graduated from the University of Twente on the classification of activities through acceleration data. Post-graduation, his career progressed at the Human Performance department at TNO, the Netherlands Organization for Applied Scientific Research. His expertise is on physiological data analysis. His work

encompasses a broad spectrum of subjects, including heart rate, respiration, as well as stress and workload, with a strong emphasis on predictive modeling through machine learning techniques.



Hilbrand Oldenhuis is Professor Digital Health at the Hanze University of Applied Sciences in Groningen, The Netherlands. In his work, he focuses on the interplay between psychology and technology. One of the core research lines in his professorship is investigating the use of wearable technologies and the application

of data science to understand psychological processes and human behavior. Furthermore, he is involved in several research projects in which the implementation of technology in health-related contexts is investigated and optimized.



Rudolf Verdaasdonk is a full professor of Health Technology at the University of Twente. In his profession as a medical physicist, he has many years of experience working with medical professionals in hospitals and medical technology companies, overseeing the route from idea/invention to implementation in healthcare. He bridges the technology push

from universities with the clinical needs from health institutes connecting the key players. He advises researchers in the preparation of clinical studies, ethical committee applications and CE certification in view of the Medical Device Regulations. One of his research interests is the validation and use of wearables for health monitoring and diagnostics.



Lisette van Gemert-Pijnen PhD is a full professor (emeritus) Persuasive Health Technology at University of Twente, The Netherlands. Her research focuses on persuasive designs to increase trust and adherence to technologies and to develop methods for implementation in practice. She founded the first Center for eHealth Research that produces the CEHRES

roadmap for eHealth. She is involved in a university-wide strategic research program to accelerate the uptake and implementation of health technologies. She is section editor Health Technology Implementation Frontiers. She published several papers and books, recently published in World Economic Forum about Privacy Enhancing Technologies to facilitate (re)use of data in healthcare.