# A Real-World Data Collection for Emotion Recognition Using Smartwatches and Lessons Learned

Elias Müller*
human-centered systems lab (h-lab)
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
elias.mueller@kit.edu

Kirsten Greiner*
human-centered systems lab (h-lab)
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
kirsten.greiner@kit.edu

Ivo Benke
human-centered systems lab (h-lab)
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
ivo.benke@kit.edu

Alexander Maedche
human-centered systems lab (h-lab)
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
alexander.maedche@kit.edu

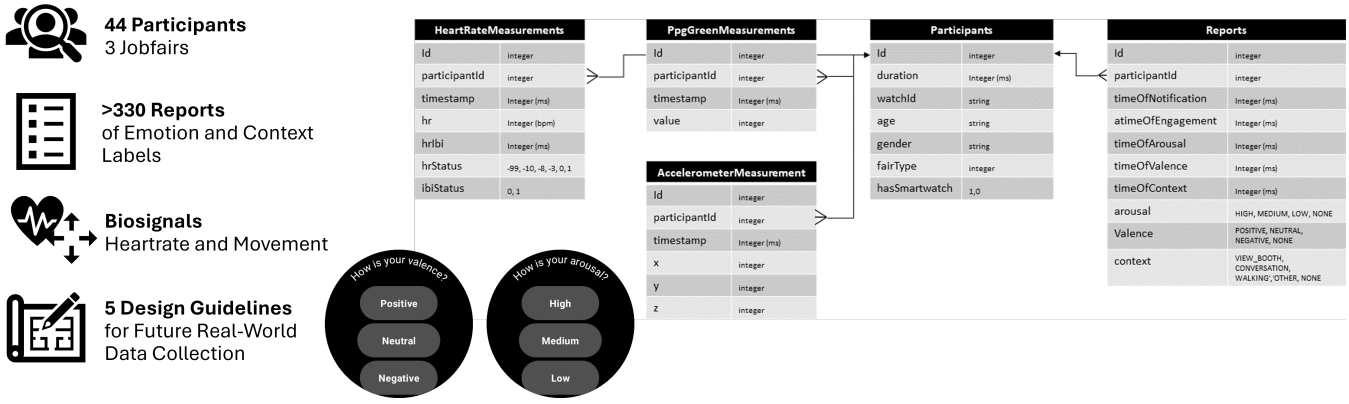



**Figure 1: Overview of the dataset, the data model, and the emotion reporting view for our data collection**

## Abstract

Emotion recognition is important across healthcare, education, and daily life, but real-world deployment remains challenging. Collecting reliable biosignal and self-report data in natural settings poses key methodological and design challenges. We introduce "BEmotion", a dataset that links optical and kinematic biosignals to self-reported valence, arousal, and context. The dataset was collected using smartwatches at three job fairs, involving 44 participants. While real-world datasets are essential, improved methods for collecting and using them are needed. Thus, we also analyze the challenges of real-world data collection encountered in the study. From these insights, we derive recommendations as well as lessons learned, and future use cases. We contribute: (1) with an open dataset and (2) recommendations for future dataset collections in the field. Furthermore, we identify a need for greater personalization in emotion recognition rather than a one-fits-all recognition.

## CCS Concepts

• **Human-centered computing → Empirical studies in ubiquitous and mobile computing**.

## Keywords

Emotion Reporting, Biosignals , Experience Sampling, Dataset

**ACM Reference Format:**
Elias Müller, Kirsten Greiner, Ivo Benke, and Alexander Maedche. 2026. A Real-World Data Collection for Emotion Recognition Using Smartwatches and Lessons Learned. In *Proceedings of Mensch und Computer 2026 (MuC '26), August 30–September 02, 2026, Duisburg, Germany*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3820253.3820282

*Both authors contributed equally to this research.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
*MuC '26, Duisburg, Germany*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2611-8/26/08
https://doi.org/10.1145/3820253.3820282

## 1 Introduction

Recognizing and processing emotions in daily life provides fundamental insights into human behavior [32, 41, 61]. Emotion recognition also enables designing emotion-adaptive systems and new forms of human–computer interaction [2, 24, 60]. Combining different types of biosignals, such as optical (e.g., photoplethysmogram)

and kinematic biosignals (e.g., accelerometer), is promising for robust emotion recognition. Significant progress has been made in developing datasets combining biosignals and self-reported emotions [39, 50]. However, real-time emotion recognition in daily life requires comprehensive datasets. These datasets should include multiple biosignals and reported emotions across diverse situations and contexts [27, 40].

Wrist-worn devices such as smartwatches are a promising approach for field data collection [12, 59]. They enable continuous, non-invasive biosignal capture while integrating into daily life [25]. Combined with Experience Sampling Method (ESM) like Ponnada et al. [46], they also support self-reported emotions and facilitate large-scale field datasets for real-time emotion recognition.

Several public datasets combine wearable biosignals and emotions, supporting emotion research [33, 35, 52, 56]. However, most were collected in controlled laboratory environments. They lack naturalistic data that reflect emotional complexity in daily life and during major events. We identified a representative set of such datasets that cover biosignals and emotions and have been collected in naturalistic settings, including the TILES [36], the K-EmoCon [42], the DAPPER [55], the MDEAW [38], the LAUREATE [28], the G-REx [5], and the WorkStress3D [14] dataset. Despite these valuable resources, more field data across contexts is needed to improve external validity and understanding of real-life emotions [27]. We therefore introduce the Biosignals and Emotions in the Field (BEmotion) dataset, a public field dataset of biosignals and self-reported emotions collected via smartwatches. It includes data from 44 participants, covering heart activity and motion as well as self-reported emotions and context. Data were collected at job fairs targeting university students across multiple contexts using a dedicated smartwatch self-report app.

Building on this dataset and the practical insights gained during its collection, our work makes two main contributions:

- **Dataset Contribution:** We present BEmotion, an open dataset of biosignals and self-reported emotions and context collected in the field at three job fairs with 44 participants. The dataset captures optical biosignals (Photoplethysmogram (PPG)) to derive Heart Rate (HR) and Inter-beat Interval (IBI) and kinematic biosignals (Accelerometer (ACC)) in combination with experience-sampled valence, arousal, and situational context self-reported data, collected via a dedicated smartwatch app.
- **Methodological Contribution:** Drawing from practical challenges encountered during our in-the-wild data collection, we derive five recommendations for future data collections in the field combining biosignal recording with experience sampling. These recommendations address signal-quality monitoring, self-report scale design, missing data handling, baseline measurement, and session duration planning.

In addition, we discuss promising use cases for future work in emotion recognition in the wild, including multimodal fusion, personalized classifiers via transfer learning, and Large-Language Model (LLM)-based in-context approaches to emotion recognition.

# 2 Related Work

## 2.1 Emotions

Emotions are a mental experience that occurs in response to specific events or experiences [17]. They involve high intensity and high hedonic content, which refers to the degree of pleasure or displeasure associated with the emotion [7]. A well-established framework to classify emotions is the two-dimensional circumplex model [49], which posits that basic emotions can be mapped along the two dimensions: valence and arousal. [47]. Valence refers to the positive or negative evaluation of an emotional state, or whether an emotion is perceived as pleasant or unpleasant [4]. Positive valence describes emotions such as happiness, joy and satisfaction, while negative valence includes emotions such as fear, sadness and anger. Arousal describes the intensity or strength of an emotion [4]. This dimension ranges from low arousal, such as relaxed or calm, to high arousal, such as excited or anxious. Each human emotion can be described as a linear combination of these two dimensions [47]. A visual depiction of the model by Russell [49] is shown in figure 2. The two-dimensional circumplex model provides a concise and comprehensive way to categorize and understand the wide range of human emotions.

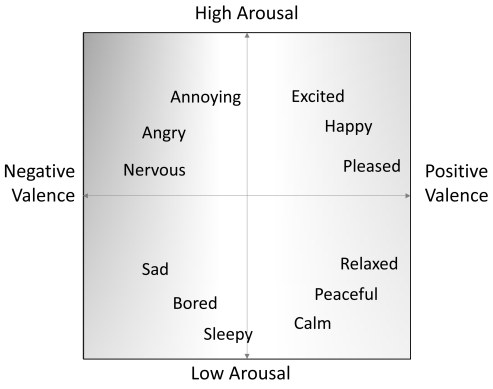



**Figure 2: Two Dimensional Emotion Model**

Emotion recognition encompasses a range of methodologies and techniques aimed at identifying and categorizing human emotions based on various cues [18], including facial expressions [47], vocal intonations [19], physiological signals [63], or behavioural patterns [43]. Machine learning algorithms are commonly employed to analyze these cues and infer the underlying emotional states [61]. Recognizing emotions in real-world settings, particularly during events and public activities, is essential for gaining insights into human behaviour [20]. Multiple researchers have studied emotion recognition in controlled experiment settings in abundant ways. Nevertheless, by studying emotions in the field, researchers can elucidate the complex interplay between individual experiences, external influences, and emotional responses [29].

## 2.2 Smartwatch-based Experience Sampling

The idea of experience sampling has a long history and builds on the research of Burtt [6]. With ESM, self-reports are obtained from

people over the course of the day and offer the possibility of experiencing intra-psychic elements of human life [57]. In particular, ESM is a diary method for systematically obtaining self-reports from people about what they do, feel and think during activities in their daily lives [10]. Different devices can be leveraged in ESM studies, e.g. pagers, pen and paper, mobile devices, wearables, or PC-based applications. Independently of the device, participants are interrupted at random points in their daily lives and asked to answer a survey capturing current experiences. Furthermore, different devices can be leveraged to collect biosignal data in combination with the self-reported surveys in experience sampling.

Mobile devices such as smartphones, wrist-worn devices, and head-mounted wearables provide new means for applying experience sampling in the field. In a comparative study by Hernandez et al. [21] wrist-worn and head-mounted wearables were compared with a conventional smartphone in a field study. Although no device stood out from the others, the response time of the wrist-worn device wearers was much lower using an ESM application with a 2D grid in comparison with the two other devices. A recently published study conducted by Schneegans et al. [53] examined the feasibility and usability of smartwatch-based experience sampling in a laboratory setting with 9 participants and reported high levels of usability. Smartwatches represent particularly promising experience-sampling devices as they enable users to seamlessly self-report through their displays and record biosignal data via their built-in sensors. For example, the Samsung Watch 5 Pro has various sensors [51], including a PPG and ACC, which we also leverage for our dataset. The signal provided by a PPG sensor can be used for optical volumetric measurement of organs [48]. Due to their widespread use and their use in smartwatches, PPG sensors can be used to collect the required raw data as a foundation to compute Heart Rate Variability (HRV) features, which in combination with experienced sampled self-reported data can be used to recognise emotions [8]. Many publications have successfully used PPG sensors as a foundation to identify emotions [30, 54]. In addition, the ACC measures a subject's acceleration or body movement, in our example, the movement of the arm on which the watch was worn. Recently, a growing interest was placed in classifying emotions from body movements. From a psychological point of view, a link between emotions and body movement is well established[9, 13, 45, 58]. Based on this, Ahmed et al. [1] proposes a system that achieves a recognition accuracy during walking of 90 %, when sitting 96 %, and in action-independent cases 86.66 %.

## 2.3 Existing Datasets

Several publicly available datasets exist that capture emotions and biosignals. Examples include the Affectiva-MIT Facial Expression Dataset (AM-FED) [34], the Database for Emotion Analysis using Physiological Signals (DEAP) [26], and the Multimodal Dataset for Wearable Stress and Affect Detection (WESAD) [52]. These datasets encompass a range of modalities, including facial expressions and physiological signals, providing valuable resources for emotion research and technology development. However, these datasets are constrained to laboratory or controlled experiments, limiting their validity and transferability to everyday contexts.

While existing datasets offer valuable insights into emotion recognition and analysis, we argue that there remains a gap regarding the availability of comprehensive datasets that capture emotions and biosignals in naturalistic, real-world settings. There is a need to provide datasets that reflect the complexity and variability of emotions experienced in the field, including daily activities, social interactions, or significant life events. We conducted an exploratory review of representative datasets focused on emotions and field-based biosignal data collection, with a particular focus on datasets that jointly provide physiological biosignals acquired through wearable and corresponding emotion or affect annotations. Table 1 provides a brief overview of selected field-collected datasets.

The TILES dataset (Tracking Individuals in Longitudinal Environmental Studies) by [36] is a longitudinal real-world dataset collected in a hospital work environment. It comprises multimodal physiological and environmental biosignals recorded using wearable and ambient sensing technologies. In addition, the dataset provides self-reported affective and well-being measures collected via repeated surveys. The dataset is available in two releases, TILES-2018 and TILES-2019. The K-EmoCon dataset (a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations) by Park et al. [42] contains biosignal data during 10-minute debates on social topics. In addition, data on valence, arousal, and 18 categorical emotions were collected through ex post annotations by the two debate partners and an external observer during the audiovisual recording of each debate. The DAPPER dataset (Daily Ambulatory Psychological and Physiological Recording for Emotion Research) by Shui et al. [55] used the experience-sampling method and the day reconstruction method to collect participants' emotional experiences over a 5-day study. Additionally, biosignal data were collected with a custom-designed wristband. The MDEAW dataset (a multimodal dataset for emotion analysis from wireless wearable devices) by Nandi et al. [38] contains Electrodermal Activity (EDA) and PPG signals recorded from students during real classroom and exam-related activities. Emotional states were annotated via self-reports of six basic emotions. At present, the dataset is available only as a preprint on arXiv. Laporte et al. [28] presents LAUREATE (a dataset for supporting research in affective computing and human memory augmentation), a longitudinal study over 13 weeks to monitor lectures and exams during one university semester. In addition to biosignal data, emotion recognition was based on physiological and behavioural changes. Bota et al. [5] present the G-REx dataset (a real-world dataset of group emotion experiences based on physiological data). The dataset contains biosignal data collected with a wrist-worn device recorded during different movie sessions in groups. In regard to emotion, the participants had to annotate selected segments of a few seconds (20 seconds) where events with high amplitude were detected for the subjects on the basis of EDA data. Finally, the WorkStress3D dataset by Dogan et al. [14] provides biosignal data collected in real workplace environments using wearable sensors. It includes EDA, Blood Volume Pulse (BVP), skin temperature, as well as self-reported stress and affect measures.

In summary, while these publicly available datasets offer valuable resources for emotion research in the field, there is a need for greater emphasis on field collection of emotions and biosignals across multiple contexts to increase external validity and advance understanding of human emotions in real-life settings. DAPPER and

**Table 1: Existing Datasets of Data Collections in the Field**

| Dataset | Year | Participants | Sensor Data | Setting |
|---|---|---|---|---|
| TILES [36] | 2018 | 212 | ECG, PPG | hospital work |
| K-EmoCon [42] | 2020 | 32 | BVP, ECG, EDA, EEG, TEMP | naturalistic conversations |
| DAPPER [55] | 2021 | 88+ | ACC, HR, TEMP | every day life |
| MDEAW [38] | 2022 | 10 | EDA, PPG | education (exam) |
| LAUREATE [28] | 2023 | 42 | ACC, EDA, HR, IBI, PPG, TEMP | education (lecture sessions) |
| G-REx [5] | 2024 | 190+ | EDA, PPG | cinema (movie sessions) |
| WorkStress3D [14] | 2024 | 20 | BVP, EDA, TEMP | work |

LAUREATE are the two datasets closest to our BEmotion dataset. In contrast to these two valuable datasets, we collect our data in a new specific context (job fairs) and use a commercially available smartwatch to integrate biosignal and self-reported emotion data collected through our self-reported experience sampling app.

# 3 Method

## 3.1 Data Collection Procedure

We chose job fairs as the study context because interactions with employers can trigger positive and negative emotions. Participants moved continuously through different situations, which we captured in a naturalistic setup. Participation was voluntary, and data were collected at three job fairs (see Table 2): Stuttgart, Heidelberg, and Karlsruhe. The same data collection procedure was used at all job fairs:

(1) We set up an information table at each fair and approached participants directly to invite them to the study. Participants received 10.00€ per hour for participation and wearing the smartwatch for biosignal collection. They additionally received 0.50€ for each automatically triggered emotion report completed every 10–15 minutes.
(2) After a brief introduction, a short demographic questionnaire, and instructions on the smartwatch and the concepts of valence and arousal, recording via the smartwatch app happened.
(3) Participants then recorded a baseline in silence while seated calmly at our table and watching a 4:54-minute calming video.
(4) After the baseline, participants freely explored the job fair and pursued their activities.
(5) During the visit, they were automatic notification triggered every 10 to 15 minutes via smartwatch to report emotions and context. This relatively short interval was chosen to capture emotional fluctuations close to the underlying experiences while reducing retrospective recall bias in self-reports. At the same time, the interval was considered long enough to minimize excessive interruption during job fair participation.
(6) After one to three hours, participants returned the watch and completed the session. The session duration was not predetermined by the study protocol but depended on how long participants naturally chose to remain at the job fair. This approach preserved ecological validity by allowing participants to engage with the fair according to their individual needs and schedules while enabling repeated emotion sampling throughout their visit.

## 3.2 Data Collection App

We developed an Android app in Android Studio using Kotlin to record raw biosignals and ESM data on emotions and context. The app was deployed on a Samsung Watch 5 Pro. Biosignal collection was implemented using the Samsung Privileged Health SDK[1]. Using this official package allowed use to collect the following biosignal: PPG in 25 Hz, HR in 1 Hz, IBI in 1 Hz and 3-axis ACC with 25 Hz.

To capture emotions and context via ESM, we implemented a periodic worker that sent notifications[2]. The worker started at the beginning of each session. Participants reported valence (high, neutral, low) and arousal (positive, neutral, negative) based on the Circumplex Model of Russell [49]. The use of a three-point response scale represents a simplified form of experience sampling, aligned with ecological momentary assessment (EMA) and micro-EMA approaches [22] that prioritize low cognitive burden and rapid in-the-moment reporting in real-world settings. While standardized instruments such as the Positive and Negative Affect Schedule (PANAS) and continuous visual analog scales are widely used in affective measurement, they typically involve higher response effort and may be less suitable for frequent mobile data collection scenarios. In addition, participants also indicated their current activity (viewing a booth, walking, conversing, or other). The first report occurred after 6 min (one minute after baseline). Subsequent reports were triggered every 10 to 15 minutes.

# 4 BEmotion: The Dataset

## 4.1 Participants

We had a total of 44 participants at all three job fairs. All participants gave their written consent for us to publish the data. The ethics committee of our institution reviewed and approved the declaration of consent, the data security and privacy policy as well as the study procedure. Table 9 shows the distribution of individual participants per job fair.

In total, 44 participants participated in the study at the three job fairs. 28 (63.6 %) were between 18 and 25 years old, 10 (22.7 %) between 26 and 30 years and six (13.6 %) between 31 and 40 years

[1] https://developer.samsung.com/health/privileged/overview.html
[2] using the Androidx Work packagehttps://developer.android.com/jetpack/androidx/releases/work sending a notification using Android Notification Manager https://developer.android.com/reference/android/app/NotificationManager

**Table 2: Overview of the three job fairs**

| # | Date | City | Participants | Context |
|---|---|---|---|---|
| 1 | 21 October 2023 | Stuttgart | 9 | No focus with openings, training, and further education opportunities; exhibitors included global players, SMEs, and educational institutions; no prior registration required |
| 2 | 25 November 2023 | Heidelberg | 13 | Focused on medical students; information on internships, specializations, and workplaces; clinics as exhibitors; workshops, lectures, and panels; free prior registration required |
| 3 | 18 January 2024 | Karlsruhe | 22 | Focused on technical, scientific and business careers; employer networking, company presentations and portfolio checks; exhibitors from various industries |

old. Of the 44 participants, 27 (61.4 %) were female and 17 (38.6 %) were male. 14 (31.8 %) of the 44 participants own a smartwatch and have corresponding experience, whereas the other 30 (68.2 %) do not. In total, all sessions of all participants together lasted 84 h. On average, the participants took part 1 h:54 min (SD: 31 min) in the study.

## 4.2 Data Model

The data collected for BEmotion is represented by three parts: (1) participant, (2) biosignals, and (3) self-reported data. Figure 4 depicts the underlying data model of our dataset. It consists of five main interconnected entities that are organized correspondingly in five Comma-Separated Values (CSV)-files in our dataset.

The dataset is organized around the participants' session entity. This includes the following core information about the participants: participant ID, watch ID, job fair ID, session duration, and basic demographic information. The time captured in the dataset is in unix epoch time in milliseconds related to the start time of the participant's session of the participant. The unix epoch time can be easily converted into a human-readable date using publicly available converters. The attributes of the corresponding CSV file contained in our dataset are described in the Appendix in Table 4.

As introduced above, we extracted various biosignal data using the Samsung SDK for the Galaxy Watch 5 Pro. First, we extracted ACC raw data from the sensor. This provides information about the movement behavior of our participants (timestamp, x-value, y-value, z-value). The attributes of the corresponding CSV file contained in our dataset are described in the Appendix in Table 5. Second, we extracted HR raw data from the PPG sensor of the smartwatch. Furthermore, we provide HR and IBI post-processed and provided by the Samsung SDK. The attributes of the corresponding CSV files contained in our dataset are described in Table 7 and Table 6.

Furthermore, as mentioned above, emotion and context self-reported data were collected through ESM with the smartwatch using our app. The structure of the data is captured in Figure 3. It includes timestamps for interactions, arousal reports, valence reports, and situational context reports.

## 4.3 Data Availability

BEmotion is available on the data mangement system of our institute RADAR4KIT [37]. Furthermore, we are happy to share the code of our ESM app for the smartwatch available on request. To use the code, however, the user needs to apply it on the Samsung developer page for the Privileged Health SDK.

# 5 Data Validity

## 5.1 Descriptive Analysis

Table 10 provides an overview of the descriptive statistics describing the self-reported data collected at the three job fairs. We successfully captured a wide range of valence and arousal states across different situational contexts.

Table 11 reports status information delivered by the smartwatch regarding the collected heart activity data. In particular, it provides an overview of the three job fairs for the HR and IBI statuses. Around 43.5 % of the recordings were classified as status −10, indicating either insufficient PPG signal quality or excessive movement by the wearer. In contrast, 55.8 % of the measurements were classified as successful In the IBI measurements, 23.2 % were classified as bad and the rest (76.8 %) as good.

## 5.2 Biosignal Preprocessing and Classification Results

To validate the data used for emotion recognition, biosignal recordings were extracted from a temporal window spanning five minutes before and five minutes after the relevant event. These signals were segmented into overlapping windows of 60 s with a step size of 10 s. Within each window, HRV and ACC features were computed using FLIRT[3] [16].

In addition, a baseline period was defined from 30 seconds after the onset of the recording to 4 minutes and 30 seconds. This baseline was processed using the same sliding window configuration (60-second window length, 10-second step size), and the same set of HRV and ACC features was extracted. From the baseline features, the mean and standard deviation were computed and used to perform z-score normalization on the full dataset, thereby ensuring subject-specific normalization and reducing inter-individual variability.

Before using the data for training, we set aside approximately 20 % of the data as test data and excluded it from the training set. To do this, we ensured that data from the same participant or the same report—depending on the training—were not included in the training set (see Results 3).

To train the models, we used the TabularPredictor from AutoGluon [15]. AutoGluon is a state-of-the-art framework that greatly

[3] https://pypi.org/project/flirt/

simplifies machine learning by automatically selecting, training, and optimizing models. TabularPredictor is specifically designed for structured tabular data and automatically handles feature engineering, model selection, and hyperparameter tuning, enabling accurate predictions. The scripts could be downloaded on **GitHub**.

In our experiments, we used the "best_quality" preset, which focuses on predictive performance through extensive ensemble training and hyperparameter optimization. The models were trained to optimize the weighted F1 score (f1_weighted) and used 5-fold bagging for the cross-validation-based ensemble method. Training was limited to 1200 s and was performed on a system with a single GPU (RTX 3090) and 48 CPU cores. Longer training sessions didn't really improve accuracy.

*Results:* Table 3 summarizes the classification performance of the trained models on the held-out test set. We grouped the data into four training sets and testing sets. We followed the separation strategies. In one case, we split so that no data from a single participant appeared in both datasets, and in another, we ensured that no data from the same emotion report was included in either the training or test set. Within each strategy, three prediction targets were examined: the simultaneous classification of the combined valence-arousal label, as well as two independent models for valence and arousal separately, and the prediction of the situational context label (*Conversation*, *View Booth*, *Walking*, *Other*; see Table 10).

When **split by participant**, the combined Valence-Arousal model (RandomForestEntr) achieves an Accuracy of 0.379 and a weighted F1-Score of 0.274. The F1-Macro is 0.274 and for valence 0.388 and for arousal 0.258. When evaluated separately by dimension, ExtraTreesGini (Valence) and CatBoost (Arousal) achieve an Accuracy of 0.317 with a weighted F1-score of 0.292; the F1-macro is 0.127, 0.398 for valence and 0.376 for arousal. Context classification (CatBoost) achieves the highest values in this group with an Accuracy of 0.418 and a weighted F1-Score of 0.385; the F1-Macro is 0.268.

When **split by report**, consistently higher values are achieved. The combined valence-arousal model (ExtraTreesEntr) achieves a Accuracy of 0.446 and a weighted F1-score of 0.421, with F1-Macro values of 0.22, 0.424 for valence and 0.457 for arousal. The overall F1 macro is 0.22. The separate models (RandomForestEntr for both dimensions) achieve an Accuracy of 0.456 and a weighted F1-Score of 0.441; a F1-Macro of 0.271, for valence 0.445 and Arousal 0.482. The context classification (CatBoost) achieves the highest performance within this group with an Accuracy of 0.552, a weighted F1-Score of 0.562, and an F1-Macro of 0.421.

In a separate experiment, we investigated the extent to which the HRV and ACC features derived from the biosignal data can be used to identify individual participants. In this experiment, the data was split by reports, while the prediction target was participant affiliation. NeuralNetFastAI achieved an Accuracy of 0.956, a weighted F1-Score of 0.958, and an F1-Macro of 0.874.

# 6 Discussion

## 6.1 Interpretation of the Data Validity

The dataset comprises 44 and contains over 84 h of biosignal data, amounting to just under 1 h and 54 min per participant. The participants were mainly from the younger and middle-aged population, since the focus was on job fairs. However, the number of reports per participant is less than 10, because not all reports could always be answered, as the participants did not have the opportunity to respond to the notification at that moment. Furthermore, analysis of the heart rate data indicates that nearly 43.5 % of PPG readings are weak due to movement artifacts. This is a known issue of PPG readings [23]. Since the data were collected in real-world settings rather than in controlled settings, movement was inherent to the process and cannot be avoided.

The results in Table 3 show that the splitting strategy used in the train-test split has a substantial impact on classification performance. When split by participant, the accuracy values for valence-arousal classification range between 0.317 and 0.379; when split by report, values of up to 0.456 are achieved. The moderate overall performance in emotion recognition clearly highlights the difficulty of generalizing from field data. The separate participant identification experiment underscores this: By predicting participant affiliation with an Accuracy of 0.956, which could suggest that the models preferentially learn person-specific patterns rather than emotion-relevant cross-person patterns.

The consistently low F1-Macro is due to the highly unbalanced class distribution: Negative valence was reported only 29 times, compared to 151 positive and 159 neutral reports (see Table 10), which results in rare classes being detected less accurately.

The context classification achieves an accuracy of 0.552 and a weighted F1 score of 0.562 when grouped by report. Context labels such as *Walking* or *Conversation* are objectively and consistently represented across individuals in the ACC signal, whereas valence and arousal are subject to a stronger subjective interpretation and thus reduce class separability.

We see strong potential for a transfer learning approach in which a base classifier pre-trained on BEmotion is fine-tuned to individual or groups, offering a promising path toward personalized emotion recognition in the wild. This personalization approach is not new for emotion recognition Li and Washington [31], Perz [44]. Saganowski et al. [50] address the so-called cold-start problem in intense emotion recognition, where no user-specific data is initially available, by proposing pre-trained models combined with group-level personalization. Such approaches require a dataset, such as BEmotion, with a sufficient amount of training data to solve the cold-start problem, lay the groundwork for personalized emotion recognition in future research, and enable emotion recognition in the real world.

## 6.2 Limitations of the Dataset

One limitation of the dataset is the context of a job fair, which may not capture all facets of human emotions and interactions. Nevertheless, we believe this context is transferable to events in general, especially when participants move around a lot. In addition, certain participant groups are over- or underrepresented in gender and age group, which can lead to distortions in the data. This is

**Table 3: Test-set performance (valence-arousal)**

| Split Criteria | Prediction | Model | Accuracy | F1 macro | F1 weighted | F1 macro Valence | F1 macro Arousal |
|---|---|---|---|---|---|---|---|
| participant | valence-arousal | RandomForestEntr | 0.3790 | 0.1031 | 0.2735 | 0.3881 | 0.2584 |
| | valence-arousal (seperate models) | ExtraTreesGini & CatBoost | 0.3172 | 0.1270 | 0.2920 | 0.3984 | 0.3760 |
| | context | CatBoost | 0.4177 | 0.2677 | 0.3847 | | |
| report | valence-arousal | ExtraTreesEntr | 0.4464 | 0.2195 | 0.4206 | 0.4239 | 0.4571 |
| | valence-arousal (seperate models) | RandomForestEntr & RandomForestEntr | 0.4556 | 0.2705 | 0.4413 | 0.4450 | 0.4822 |
| | context | CatBoost | 0.5519 | 0.4205 | 0.5619 | | |
| report | participant | NeuralNetFastAI | 0.9556 | 0.8736 | 0.9576 | | |

due to the way we recruited our participants. We approached them directly at the job fairs, without pre-screening, as it was important to us that our participants were genuinely interested in attending the fair and, therefore, likely to be more emotional during their attendance.

A further limitation of the data set could be that the participants understood the terms valence and arousal in the self-reports differently and incorrectly, despite a detailed introduction at the beginning of each session. The prevalence of many neutral emotions in the self-reports raises questions about the reliability and accuracy of participants' ratings. This may be due to several factors, including the complexity of emotional experiences and participants' subjective interpretations of the valence and arousal dimensions. We tried to counteract the misunderstanding of valence and arousal by ensuring that each participant had a good understanding of the terms at the beginning of each session. We achieved this through clear definitions and some exemplary categorizations according to our self-report scheme on the smartwatch for different emotional states. There is a possibility of bias in self-reporting due to participants' limited awareness or misunderstanding of their own emotional states. Despite efforts to provide comprehensive definitions of valence and arousal, personal subjectivity remains a problem. The decision to abandon the selection of discrete emotions and instead ask for valence and arousal may have an impact on the depth of emotional insights that can be gained from the self-reported data. However, this decision aimed to simplify the evaluation process and improve the user experience.

For the collected biosignals, we opted for HR/IBI and ACC data due to their proven relevance and accuracy for emotion recognition. However, the challenges associated with biosignal measurement, particularly the inaccuracies of the PPG sensors, require careful consideration. The integration of the HR/IBI data calculated by the smartwatch mitigated these inaccuracies and improved the biosignal analysis's robustness, but on the other side, this processing step is done by the manufacturer of the watch and is not documented transparently, which is why the PPG raw data is included. Investigating further biosignals, such as skin temperature or EDA, posed a challenge in the implementation of the current study. Future research could also include these biosignals. In our view, the focus on HR/IBI and ACC data represents an understandable decision for our setting, as they are appropriate for emotion recognition and the context of a job fair.

## 6.3 Lessons Learned from Real-World Data Collection

The analysis of the BEmotion dataset has yielded several methodological insights and informed a set of recommendations that can support the planning and implementation of future field studies combining biosignal recording and experience sampling.

*Watch Fit and Signal Quality Monitoring.* Although participants were instructed at the beginning of each session to wear the smartwatch snugly, the correct fit was not monitored further throughout the study. As a result, approximately 43.5 % of the HR measurements were classified as status −10, indicating insufficient PPG signal quality due to excessive movement or poor sensor contact. Future studies should consider integrating real-time signal-quality feedback directly into the smartwatch application, for example, by alerting participants when the HR status remains at −10 for an extended period. Such an in-app warning would discreetly prompt participants to reposition the watch without requiring intervention by the researchers, which is particularly advantageous in unsupervised field studies. This highlights the importance of integrating real-time signal quality feedback into wearable systems to ensure robust biosignal collection in the wild.

*Continuous Scale for Self-Reported Emotions.* Valence and arousal were measured using three discrete options (positive, neutral, negative; or high, medium, low). In hindsight, a continuous slider-based input could provide a more fine-grained response resolution, although this does not necessarily imply greater psychological granularity in the underlying emotional state. Discrete scales with a neutral midpoint tend to skew responses toward the midpoint, whereas continuous scales reduce this tendency and can capture finer emotional states. However, both measurement formats introduce different types of response bias and do not directly capture the complexity of emotional experiences. Although a slider on a small smartwatch display makes usability difficult, future work should explore interaction designs that balance input precision with ease of use in mobile contexts. This suggests prioritizing continuous emotion scales over discrete categories, while ensuring that their interaction design remains usable on small wearable displays.

*Handling Missing Self-Reports.* A significant number of self-reports yielded *None* as responses, often because participants were engaged in conversations at the time of the notification, precisely those moments that are particularly emotionally relevant. To reduce missing reports, future studies could implement a brief follow-up notification two to three minutes after an unanswered prompt, thereby giving participants a second opportunity to respond without significantly increasing the frequency of interruptions. More intrusive measures, such as louder or longer notifications, could further increase the response rate but would contradict the study's naturalistic design goals. Since BEmotion was specifically designed to capture emotions under everyday field conditions, a certain proportion. This highlights the importance of balancing response rates and intrusiveness by using subtle follow-up mechanisms rather than more disruptive notification strategies.

*Setting and Stimulus for the Baseline Measurement.* The baseline was recorded using a 4 min 54 s calming video while participants sat at the study table in the open testing environment. In retrospect, two learnings became apparent. First, the video stimulus was not entirely emotionally neutral and may have elicited mild affective reactions, thereby compromising the validity of the baseline. A static, visually neutral image or a uniform gray screen would have ensured a better-controlled and more reproducible resting state. Second, the baseline was conducted amid the ambient noise and social context of the job fair, which likely increased participants' physiological arousal relative to a quiet environment. A quiet side room at the fair would have been sufficient and methodologically more appropriate. At the same time, conducting the baseline under naturalistic conditions reflects the real-world circumstances under which the subsequent measurements also took place, which is consistent with the goals of ecological validity in a field study. This highlights the importance of combining neutral baseline stimuli with sufficiently controlled environments, while still aligning with the ecological context of the study.

*Session Duration and Longitudinal Depth.* Participants participated in the study for an average of approximately 2 hours. Although this yielded a rich cross-sectional dataset across 44 participants, future research could complement this design by measuring fewer participants over a full day (8 or more hours). Longer sessions would capture a broader spectrum of emotional states and situational contexts within individual participants, thereby enabling more robust within-person analyses and supporting the development of personalized emotion recognition models. This highlights the importance of balancing sample size and session length depending on whether the focus lies on cross-sectional insights or within-person dynamics.

*Derived Recommendations for Future Data Collections with Biosignals in the Field.* Building on the findings from the BEmotion data collection, we identified a set of recommendations intended to support future studies that combine biosignal recording with the collection of user experiences in real-world contexts. These recommendations go beyond the specific design of our study and aim to provide researchers and practitioners with actionable guidance.

- **Provide real-time signal quality feedback** to enable participants to maintain proper sensor contact during data collection.
- **Use continuous self-report scales** to capture nuanced emotional states, while ensuring usability on wearable devices.
- **Incorporate low-intrusive follow-up prompts** to reduce missing self-reports without compromising ecological validity.
- **Design neutral and controlled baseline conditions** while carefully balancing experimental control with real-world context.
- **Align session duration with analysis goals**, prioritizing longer recordings for within-person insights and personalization.

Together, these recommendations highlight the importance of balancing data quality, user burden, and ecological validity when conducting field studies.

## 6.4 Use-Cases of the Dataset

Although our work is subject to the limitations mentioned above, this dataset holds potential for application, as the data were collected not in a laboratory setting but in participants' real-world situations and could be used in similar settings. We present several use cases that could serve as a foundation for future work.

*Emotion Recognition in the Wild.* This dataset could be used to train models that are robust to motion artifacts and variable field conditions. Datasets recorded in the laboratory can be used and supplemented with data from BEmotion, thereby enabling models to operate outside of controlled settings. In addition, synthetic data generators could be explored to augment the dataset for training.

*Multimodal Emotion Recognition in the Wild.* The multimodal sensor approach of BEmotion enables researchers to compare movement (ACC) and heart (HR,IBI) data and calculate feature importance across modalities. It also enables researchers to examine the extent to which movement data can assist in emotion recognition when heart data are of poor quality, using movement features derived from wrist-worn sensors such as gait or gestures. Like natural walking Cui et al. [11], Zhang et al. [62] or gestures characteristics could be related to emotions via nonverbal communication [3].

*Cold Start Dataset for Personalized Emotion Recognition for Everyday Life.* With 44 participants and individual sessions, it is possible to build personalized emotion classifiers or use the data to inform personalized classifiers. Given the given amount of data, BEmotion can be used to train a base classifier, which can then be fine-tuned with individual-specific data to address cold-start issues.

*Emotion Recognition with In-Context Learning (ICL) using LLMs.* The structured ESM reports (valence, arousal, context, timestamp) lend themselves well to being formatted as few-shot examples. Researchers could prompt LLMs or multimodal models with a few labeled time windows per person and explore whether ICL works for emotion recognition in real-world settings.

*Context Detection on Similar Events.* Context labels (viewing a booth, walking, talking, other) combined with ACC data enable

the detection of context in job-fair-like events. For example, this data can also be applied to other fairs to determine the context in which a person is, and to enable the possibility of adapting the system. Furthermore, continuous ACC recordings, combined with contextual labels, provide opportunities to test the robustness of smartwatch-based gesture recognition systems, for example, by identifying false positives caused by natural arm and body movements during everyday interactions.

## 7 Conclusion

Existing datasets provide valuable insights into emotion recognition and analysis. However, capturing emotions and biosignals in naturalistic environments remains challenging, as most existing datasets are collected under controlled conditions. In this paper, we presented BEmotion, a dataset of biosignals and emotions using smartwatches in the field. The data was collected at three different job fairs, with 44 participants reporting their emotions and simultaneously recording their biosignals. The dataset captures heart activity (HR, PPG) as well as motion data (ACC). Furthermore, emotion (valence, arousal) and context (view booth, walking, conversation, other) reports were collected using a dedicated self-reporting app we developed and provided on the smartwatch. Beyond describing the dataset and data collection pipeline, we reflected on the practical challenges encountered during real-world data collection. From these experiences, we derived methodological insights, formulated recommendations for future data collections with biosignals in the field, and outlined potential use cases for BEmotion. Together, these contributions establish BEmotion not only as an open dataset, but also as a methodological foundation for future emotion recognition research in real-world contexts.

## Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) -GRK2739/2-Project Nr. 447089431 - Research Training Group: KD2School - Designing Biosignal-Adaptive Systems for Decision-Making Processes.

## References

[1] Ferdous Ahmed, A. S. M. Hossain Bari, and Marina L. Gavrilova. 2020. Emotion Recognition From Body Movement. *IEEE Access* 8 (2020), 11761–11781. doi:10.1109/ACCESS.2019.2963113

[2] Mina Alipour, Mahyar Tourchi Moghaddam, Karthik Vaidhyanathan, and Mikkel Baun Kjærgaard. 2023. Toward Changing Users behavior with Emotion-based Adaptive Systems. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. ACM, Limassol Cyprus, 85–95. doi:10.1145/3565472.3595614

[3] Esma Nur Asalıoğlu and Tilbe Göksun. 2023. The role of hand gestures in emotion communication: Do type and size of gestures matter? *Psychological Research* 87, 6 (Sept. 2023), 1880–1898. doi:10.1007/s00426-022-01774-9

[4] Saikat Basu, Nabakumar Jana, Arnab Bag, Mahadevappa M, Jayanta Mukherjee, Somesh Kumar, and Rajlakshmi Guha. 2015. Emotion recognition based on physiological signals using valence-arousal model. In *2015 Third International Conference on Image Information Processing (ICIIP)*. IEEE, Waknaghat, India, 50–55. doi:10.1109/ICIIP.2015.7414739

[5] Patrícia Bota, Joana Brito, Ana Fred, Pablo Cesar, and Hugo Silva. 2024. A real-world dataset of group emotion experiences based on physiological data. *Scientific Data* 11, 1 (Jan. 2024), 116. doi:10.1038/s41597-023-02905-6

[6] Harold E. Burtt. 1932. Review of Workers' Emotions in Shop and Home. *Journal of Applied Psychology* 16, 5 (1932), 592–594. doi:10.1037/h0068164 Place: US.

[7] Michel Cabanac. 2002. What is emotion? *Behavioural processes* 60, 2 (2002), 69–83. doi:10.1016/S0376-6357(02)00078-5

[8] Gloria Cosoli, Angelica Poli, Lorenzo Scalise, and Susanna Spinsante. 2021. Heart Rate Variability Analysis With Wearable Devices: Influence of Artifact Correction Method on Classification Accuracy for Emotion Recognition. In *2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, Glasgow, United Kingdom, 1–6. doi:10.1109/I2MTC50364.2021.9459828

[9] Mark Coulson. 2004. Attributing Emotion to Static Body Postures: Recognition Accuracy, Confusions, and Viewpoint Dependence. *Journal of Nonverbal Behavior* 28, 2 (June 2004), 117–139. doi:10.1023/B:JONB.0000023655.25550.be

[10] Mihaly Csikszentmihalyi. 2014. *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*. Springer Netherlands, Dordrecht. doi:10.1007/978-94-017-9088-8

[11] Liqing Cui, Shun Li, and Tingshao Zhu. 2016. Emotion Detection from Natural Walking. In *Human Centered Computing*, Qiaohong Zu and Bo Hu (Eds.). Springer International Publishing, Cham, 23–33. doi:10.1007/978-3-319-31854-7_3

[12] Ruixuan Dai, Chenyang Lu, Linda Yun, Eric Lenze, Michael Avidan, and Thomas Kannampallil. 2021. Comparing stress prediction models using smartwatch physiological signals and participant self-reports. *Computer Methods and Programs in Biomedicine* 208 (Sept. 2021), 106207. doi:10.1016/j.cmpb.2021.106207

[13] Winand H Dittrich, Tom Troscianko, Stephen E G Lea, and Dawn Morgan. 1996. Perception of Emotion from Dynamic Point-Light Displays Represented in Dance. *Perception* 25, 6 (June 1996), 727–738. doi:10.1068/p250727

[14] Gulin Dogan, Fatma Patlar Akbulut, and Cagatay Catal. 2024. Biosignals, facial expressions, and speech as measures of workplace stress: Workstress3d dataset. *Data in Brief* 54 (June 2024), 110303. doi:10.1016/j.dib.2024.110303

[15] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. *arXiv preprint arXiv:2003.06505* (2020).

[16] Simon Föll, Martin Maritsch, Federica Spinola, Varun Mishra, Filipe Barata, Tobias Kowatsch, Elgar Fleisch, and Felix Wortmann. 2021. FLIRT: A Feature Generation Toolkit for Wearable Data. *Computer Methods and Programs in Biomedicine* (2021). doi:10.1016/j.cmpb.2021.106461

[17] James J. Gross. 2015. Emotion Regulation: Current Status and Future Prospects. *Psychological Inquiry* 26, 1 (Jan. 2015), 1–26. doi:10.1080/1047840X.2014.940781

[18] Hatice Gunes and Maja Pantic. 2010. Automatic, Dimensional and Continuous Emotion Recognition:. *International Journal of Synthetic Emotions* 1, 1 (Jan. 2010), 68–99. doi:10.4018/jse.2010101605

[19] Wen-Jing Han, Hai-Feng Li, H Bin Ruan, and Lin Ma. 2014. Review on speech emotion recognition. *Journal of software* 25, 1 (2014), 37–50.

[20] Judith S. Heinisch, Christoph Anderson, and Klaus David. 2019. Angry or Climbing Stairs? Towards Physiological Emotion Recognition in the Wild. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, Kyoto, Japan, 486–491. doi:10.1109/PERCOMW.2019.8730725

[21] Javier Hernandez, Daniel McDuff, Christian Infante, Pattie Maes, Karen Quigley, and Rosalind Picard. 2016. Wearable ESM: differences in the experience sampling method across wearable devices. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, Florence Italy, 195–205. doi:10.1145/2935334.2935340

[22] Stephen Intille, Caitlin Haynes, Dharam Maniar, Aditya Ponnada, and Justin Manjourides. 2016. µEMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 1124–1128. doi:10.1145/2971648.2971717

[23] Shahid Ismail, Usman Akram, and Imran Siddiqi. 2021. Heart rate tracking in photoplethysmography signals affected by motion artifacts: a review. *EURASIP Journal on Advances in Signal Processing* 2021, 1 (Jan. 2021), 5. doi:10.1186/s13634-020-00714-2

[24] Noel Jaymon, Sushma Nagdeote, Aayush Yadav, and Ryan Rodrigues. 2021. Real Time Emotion Detection Using Deep Learning. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*. IEEE, Bhilai, India, 1–7. doi:10.1109/ICAECT49130.2021.9392584

[25] Matin Kheirkhahan, Sanjay Nair, Anis Davoudi, Parisa Rashidi, Amal A. Wanigatunga, Duane B. Corbett, Tonatiuh Mendoza, Todd M. Manini, and Sanjay Ranka. 2019. A smartwatch-based framework for real-time and online assessment and mobility monitoring. *Journal of Biomedical Informatics* 89 (Jan. 2019), 29–40. doi:10.1016/j.jbi.2018.11.003

[26] S. Koelstra, C. Muhl, M. Soleymani, Jong-Seok Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. 2012. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing* 3, 1 (Jan. 2012), 18–31. doi:10.1109/T-AFFC.2011.15

[27] Ronak Kosti, Jose Alvarez, Adria Recasens, and Agata Lapedriza. 2019. Context Based Emotion Recognition using EMOTIC Dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), 1–1. doi:10.1109/TPAMI.2019.2916866

[28] Matias Laporte, Martin Gjoreski, and Marc Langheinrich. 2023. LAUREATE: A Dataset for Supporting Research in Affective Computing and Human Memory Augmentation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (Sept. 2023), 1–41. doi:10.1145/3610892

[29] Fanny Larradet, Radoslaw Niewiadomski, Giacinto Barresi, Darwin G. Caldwell, and Leonardo S. Mattos. 2020. Toward Emotion Recognition From Physiological Signals in the Wild: Approaching the Methodological Issues in Real-Life Data Collection. *Frontiers in Psychology* 11 (July 2020), 1111. doi:10.3389/fpsyg.2020.01111

[30] Min Seop Lee, Yun Kyu Lee, Dong Sung Pae, Myo Taeg Lim, Dong Won Kim, and Tae Koo Kang. 2019. Fast Emotion Recognition Based on Single Pulse PPG Signal with Convolutional Neural Network. *Applied Sciences* 9, 16 (Aug. 2019), 3355. doi:10.3390/app9163355

[31] Joe Li and Peter Washington. 2024. A Comparison of Personalized and Generalized Approaches to Emotion Recognition Using Consumer Wearable Devices: Machine Learning Study. *JMIR AI* 3, 1 (May 2024), e52171. doi:10.2196/52171 Company: JMIR AI Distributor: JMIR AI Institution: JMIR AI Label: JMIR AI.

[32] Vincenzo Loia and Sabrina Senatore. 2014. A fuzzy-oriented sentic analysis to capture the human emotion in Web-based content. *Knowledge-Based Systems* 58 (March 2014), 75–85. doi:10.1016/j.knosys.2013.09.024

[33] Valentina Markova, Todor Ganchev, and Kalin Kalinkov. 2019. CLAS: A Database for Cognitive Load, Affect and Stress Recognition. In *2019 International Conference on Biomedical Innovations and Applications (BIA)*. IEEE, Varna, Bulgaria, 1–4. doi:10.1109/BIA48344.2019.8967457

[34] Daniel McDuff, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. 2013. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 881–888.

[35] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. 2021. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *IEEE Transactions on Affective Computing* 12, 2 (April 2021), 479–493. doi:10.1109/TAFFC.2018.2884461

[36] Karel Mundnich, Brandon M. Booth, Michelle L'Hommedieu, Tiantian Feng, Benjamin Girault, Justin L'Hommedieu, Mackenzie Wildman, Sophia Skaaden, Amrutha Nadarajan, Jennifer L. Villatte, Tiago H. Falk, Kristina Lerman, Emilio Ferrara, and Shrikanth Narayanan. 2020. TILES-2018, a longitudinal physiologic and behavioral data set of hospital workers. *Scientific Data* 7, 1 (Oct. 2020), 354. doi:10.1038/s41597-020-00655-3

[37] Elias Müller, Kirsten Greiner, Ivo Benke, and Alexander Maedche. 2026. BEmotion: A Real-World Data Collection for Emotion Recognition Using Smartwatches and Lessons Learned. doi:10.35097/9C5YVPN5YQYCD23S Artwork Size: 466,5 MB Pages: 466,5 MB.

[38] Arijit Nandi, Fatos Xhafa, Laia Subirats, and Santi Fort. 2022. MDEAW: A Multimodal Dataset for Emotion Analysis through EDA and PPG signals from wireless wearable low-cost off-the-shelf Devices. doi:10.48550/ARXIV.2207.06410 Version Number: 1.

[39] Wang Kay Ngai, Haoran Xie, Di Zou, and Kee-Lee Chou. 2022. Emotion recognition based on convolutional neural networks and heterogeneous bio-signal data sources. *Information Fusion* 77 (Jan. 2022), 107–117. doi:10.1016/j.inffus.2021.07.007

[40] Desmond C. Ong, Zhengxuan Wu, Tan Zhi-Xuan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki. 2021. Modeling Emotion in Complex Stories: The Stanford Emotional Narratives Dataset. *IEEE Transactions on Affective Computing* 12, 3 (July 2021), 579–594. doi:10.1109/TAFFC.2019.2955949

[41] Shantanu Pal, Subhas Mukhopadhyay, and Nagender Suryadevara. 2021. Development and Progress in Sensors and Technologies for Human Emotion Recognition. *Sensors* 21, 16 (Aug. 2021), 5554. doi:10.3390/s21165554

[42] Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. 2020. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data* 7, 1 (Sept. 2020), 293. doi:10.1038/s41597-020-00630-y

[43] Amol Patwardhan. 2017. Three-Dimensional, Kinematic, Human Behavioral Pattern-Based Features for Multimodal Emotion Recognition. *Multimodal Technologies and Interaction* 1, 3 (Sept. 2017), 19. doi:10.3390/mti1030019

[44] Bartosz Perz. 2022. Personalization of Emotion Recognition for Everyday Life Using Physiological Signals From Wearables. In *2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, Nara, Japan, 1–5. doi:10.1109/ACIIW57231.2022.10086031

[45] Frank E Pollick, Helena M Paterson, Armin Bruderlin, and Anthony J Sanford. 2001. Perceiving affect from arm movement. *Cognition* 82, 2 (Dec. 2001), B51–B61. doi:10.1016/S0010-0277(01)00147-0

[46] Aditya Ponnada, Caitlin Haynes, Dharam Maniar, Justin Manjourides, and Stephen Intille. 2017. Microinteraction Ecological Momentary Assessment Response Rates: Effect of Microinteractions or the Smartwatch? *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (Sept. 2017), 1–16. doi:10.1145/3130957

[47] Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology* 17, 3 (2005), 715–734. doi:10.1017/S0954579405050340

[48] Raj Rakshit, V Ramu Reddy, and Parijat Deshpande. 2016. Emotion detection and recognition using HRV features derived from photoplethysmogram signals. In *Proceedings of the 2nd workshop on Emotion Representations and Modelling for Companion Systems*. ACM, Tokyo Japan, 1–6. doi:10.1145/3009960.3009962

[49] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161. doi:10.1037/h0077714

[50] Stanisław Saganowski, Joanna Komoszyńska, Maciej Behnke, Bartosz Perz, Dominika Kunc, Bartłomiej Klich, Łukasz D Kaczmarek, and Przemysław Kazienko. 2022. Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables. *Scientific data* 9, 1 (2022), 158. doi:doi.org/10.1038/s41597-022-01262-0

[51] Samsung Electronics GmbH. [n. d.]. Galaxy Watch5 Pro 45mm LTE \textbar Samsung DE. https://www.samsung.com/de/watches/galaxy-watch/galaxy-watch5-pro-gray-titanium-lte-sm-r925fztadbt/

[52] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM, Boulder CO USA, 400–408. doi:10.1145/3242969.3242985

[53] Tim Schneegans, Andry Niclas, Kirsten Greiner, Ivo Benke, Alexander Mädche, and Michael Beigl. 2023. Annotating Affect in the Field: A Case Study on the Usability of a Minimalist Smartwatch User Interface for Affect Annotation. In *Proceedings of the 2nd Empathy-Centric Design Workshop*. ACM, Hamburg Germany, 1–5. doi:10.1145/3588967.3588977

[54] Lin Shu, Yang Yu, Wenzhuo Chen, Haoqiang Hua, Qin Li, Jianxiu Jin, and Xiangmin Xu. 2020. Wearable Emotion Recognition Using Heart Rate Data from a Smart Bracelet. *Sensors* 20, 3 (Jan. 2020), 718. doi:10.3390/s20030718

[55] Xinyu Shui, Mi Zhang, Zhuoran Li, Xin Hu, Fei Wang, and Dan Zhang. 2021. A dataset of daily ambulatory psychological and physiological recording for emotion research. *Scientific data* 8, 1 (2021), 161. doi:10.6084/m9.figshare.13803185

[56] Luma Tabbaa, Ryan Searle, Saber Mirzaee Bafti, Md Moinul Hossain, Jittrapol Intarasisrisawat, Maxine Glancy, and Chee Siang Ang. 2021. VREED: Virtual Reality Emotion Recognition Dataset Using Eye Tracking & Physiological Measures. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (Dec. 2021), 1–20. doi:10.1145/3495002

[57] Niels van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The Experience Sampling Method on Mobile Devices. *Comput. Surveys* 50, 6 (2017), 93:1–93:40. doi:10.1145/3123988

[58] Harald G. Wallbott. 1998. Bodily expression of emotion. *European Journal of Social Psychology* 28, 6 (1998), 879–896. doi:10.1002/(SICI)1099-0992(1998110)28:6<879::AID-EJSP901>3.0.CO;2-W _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291099-0992%281998110%2928%3A6%3C879%3A%3AAID-EJSP901%3E3.0.CO%3B2-W.

[59] Gary M. Weiss, Jessica L. Timko, Catherine M. Gallagher, Kenichi Yoneda, and Andrew J. Schreiber. 2016. Smartwatch-based activity recognition: A machine learning approach. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, Las Vegas, NV, USA, 426–429. doi:10.1109/BHI.2016.7455925

[60] Haris Zacharatos, Christos Gatzoulis, and Yiorgos L. Chrysanthou. 2014. Automatic Emotion Recognition Based on Body Movement Analysis: A Survey. *IEEE Computer Graphics and Applications* 34, 6 (Nov. 2014), 35–45. doi:10.1109/MCG.2014.106

[61] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. 2020. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion* 59 (July 2020), 103–126. doi:10.1016/j.inffus.2020.01.011

[62] Zhan Zhang, Yufei Song, Liqing Cui, Xiaoqian Liu, and Tingshao Zhu. 2016. Emotion recognition based on customized smart bracelet with built-in accelerometer. *PeerJ* 4 (July 2016), e2258. doi:10.7717/peerj.2258

[63] Sicheng Zhao, Guiguang Ding, Jungong Han, and Yue Gao. 2018. Personality-Aware Personalized Emotion Recognition from Physiological Signals.. In *IJCAI*. 1660–1667.

# A Self-Report

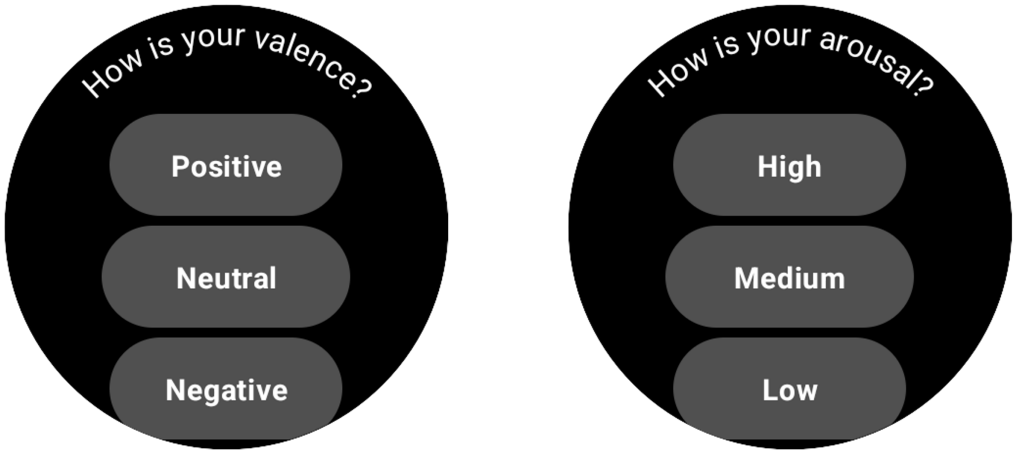



**Figure 3: Screenshot of Valence and Arousal Self-Report on Smartwatch**

## B Data Model

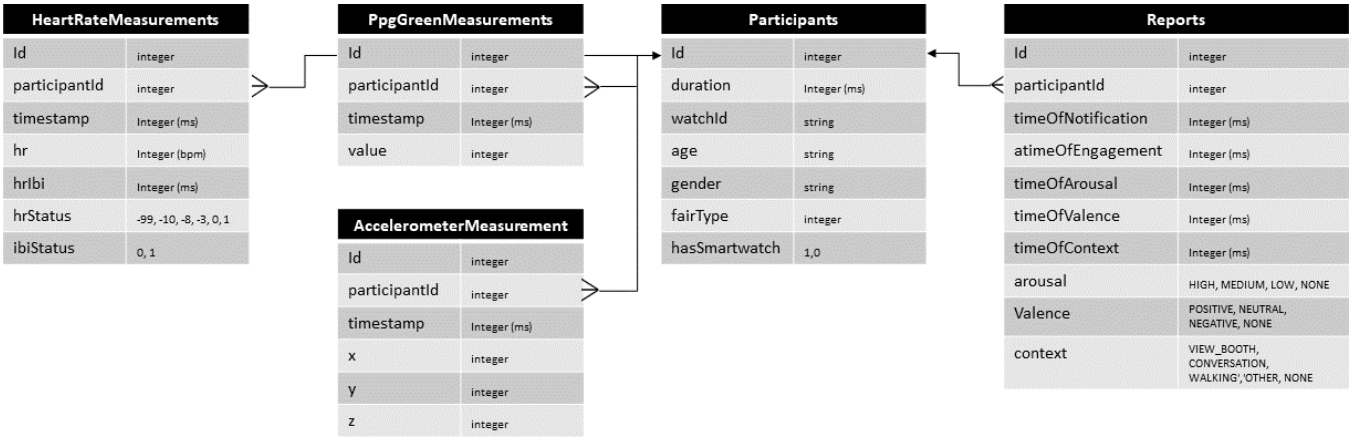



**Figure 4: Entity Relationship Diagram (ERD) of the Dataset**

**Table 4: Participant Data**

| Attribute | Description |
|---|---|
| ID | Unique participant ID |
| Duration | Duration of recorded session per participant in milliseconds |
| Watch ID | Watch ID of the smartwatch the participant wore throughout the study |
| Age | Age of the participant |
| Gender | Gender of the participant |
| Fair Type | Number of the fair (I, II, III) |

**Table 5: Accelerometer Measurement**

| Attribute | Description |
|---|---|
| ID | Unique AccelerometerMeasurement ID |
| Participant ID | Related Participant ID |
| Timestamp | Relative time reference of the measurement in milliseconds |
| X | Accelerometer's X axis value |
| Y | Accelerometer's Y axis value |
| Z | Accelerometer's Z axis value |

**Table 6: PpgGreen Measurement**

| Attribute | Description |
|---|---|
| ID | Unique PpgGreenMeasurement ID |
| Participant ID | Related Participant ID |
| Timestamp | Relative time reference of the measurement in milliseconds |
| Value | PPG green LED's ADC (analog-to-digital) value |

**Table 7: Heartrate Measurement**

| Attribute | Description |
|---|---|
| ID | Unique HeartrateMeasurement ID |
| Participant ID | Related Participant ID |
| Timestamp | Relative time reference of the measurement in milliseconds |
| HR | Heart rate value in beats per minute |
| HRIBI | IBI (inter-beat interval) in milliseconds |
| HR Status | Status flag for Heart Rate measurement. The HR Status can have the following values: [(-10)] PPG signal is too weak. Or the user's movement is too much, [(-8)] PPG signal is weak. Or there is a user's movement, [(-3)] Wearable is detached. [(0)] Initial heart rate measuring state, [(1)] Successful heart rate measurement |
| IBI Status | Status flag for IBI (inter-beat interval) (1: bad, 0: good) |

**Table 8: Self Reports**

| Attribute | Description |
|---|---|
| ID | Unique Measurement ID |
| Participant ID | Related Participant ID |
| TimeOfNotification | Relative time reference of notification to the subject |
| TimeOfEngagement | Relative time reference of the subject opening the smartwatch app in milliseconds |
| TimeOfValence | Relative time reference of the subject choosing an affect label in milliseconds |
| TimeOfArousal | Relative time reference of the subject choosing an affect label in milliseconds |
| TimeofContext | Relative time reference of the subject choosing a context label in milliseconds |
| Valence | Valence label of the subject (if the value is "NONE", the subject dismisses the label report) This field was an enumeration field, and the options were positive, neutral, or negative |
| Arousal | Arousal label of the subject (if the value is "NONE", the subject dismisses the label report) This field was an enumeration field, and the options were high, medium, or low |
| Context | Context label of the subject (if the value is "NONE", the subject dismisses the label report) |

# C Participants Demographics

**Table 9: Demographic data and metadata of the participants in the data set.**

| | | Jobfair | | |
|---|---|---|---|---|
| | **Total** | **1** | **2** | **3** |
| Location | | Stuttgart | Heidelberg | Karlsruhe |
| Participants | 44 (100.0 %) | 9 (20.5 %) | 13 (29.5 %) | 22 (50.0 %) |
| Age | | | | |
| 18-25 | 28 (63.6 %) | 7 (77.8 %) | 4 (30.8 %) | 17 (77.3 %) |
| 26-30 | 10 (22.7 %) | 0 (0.0 %) | 5 (38.5 %) | 5 (22.7 %) |
| 31-40 | 6 (13.6 %) | 2 (22.2 %) | 4 (30.8 %) | 0 (0.0 %) |
| Gender | | | | |
| Female | 27 (61.4 %) | 8 (88.9 %) | 7 (53.8 %) | 12 (54.5 %) |
| Male | 17 (38.6 %) | 1 (11.1 %) | 6 (46.2 %) | 10 (45.5 %) |
| Owns a Smartwatch | | | | |
| Yes | 14 (31.8 %) | 3 (33.3 %) | 3 (23.1 %) | 8 (36.4 %) |
| No | 30 (68.2 %) | 6 (66.7 %) | 10 (76.9 %) | 14 (63.6 %) |
| Duration | | | | |
| Sum hh:mm:ss.ms | 84:07:08.148 | 13:17:52.667 | 22:55:35.924 | 47:53:39.557 |
| Mean hh:mm:ss.ms | 01:54:42.458 | 01:28:39.185 | 01:45:48.917 | 02:10:37.252 |
| SD hh:mm:ss.ms | 00:31:46.806 | 00:29:42.294 | 00:33:41.932 | 00:20:23.732 |

# D Data Validity

**Table 10: Descriptive Statistics of the self-reported data.**

| | | Jobfair | | |
|---|---|---|---|---|
| | **Total** | **1** | **2** | **3** |
| Situational Context | | | | |
| Conversation | 112 | 22 | 11 | 79 |
| View Booth | 65 | 6 | 19 | 40 |
| Walking | 59 | 9 | 7 | 43 |
| Other | 94 | 5 | 50 | 39 |
| None | 132 | 31 | 36 | 65 |
| Valence | | | | |
| Positive | 151 | 16 | 29 | 106 |
| Neutral | 159 | 20 | 53 | 86 |
| Negative | 29 | 7 | 9 | 13 |
| None | 123 | 30 | 32 | 61 |
| Arousal | | | | |
| High | 79 | 8 | 15 | 56 |
| Medium | 191 | 22 | 48 | 121 |
| Low | 67 | 13 | 26 | 28 |
| None | 125 | 30 | 34 | 61 |

**Table 11: Overview on HR and IBI Status of the collected data.**

| | | | Jobfair | | |
|---|---|---|---|---|---|
| | **Description** | **Total** | **1** | **2** | **3** |
| HR Status | | | | | |
| -10 | PPG signal is too weak / User movement is too much | 130 923 | 31 336 | 20 835 | 78 752 |
| -3 | Wearable is detached | 889 | 94 | 367 | 428 |
| 0 | Initial heart rate measuring state | 1302 | 267 | 329 | 706 |
| 1 | Successful heart rate measurement | 167 990 | 15 911 | 60 499 | 91 580 |
| IBI Status | | | | | |
| 0 | good | 231 313 | 39 531 | 54 984 | 136 798 |
| 1 | bad | 69 817 | 8077 | 27 059 | 34 681 |