Stress Annotations from Older Adults - Exploring the Foundations for Mobile ML-Based Health Assistance

Michael Dietz
Augsburg University
Augsburg, Germany
dietz@hcm-lab.de

Ilhan Aslan
Augsburg University
Augsburg, Germany
aslan@hcm-lab.de

Dominik Schiller
Augsburg University
Augsburg, Germany
schiller@hcm-lab.de

Simon Flutura
Augsburg University
Augsburg, Germany
flutura@hcm-lab.de

Anika Steinert
Charité -
Universitätsmedizin Berlin
Berlin, Germany
anika.steinert@charite.de

Robert Klebbe
Charité -
Universitätsmedizin Berlin
Berlin, Germany
robert.klebbe@charite.de

Elisabeth André
Augsburg University
Augsburg, Germany
andre@hcm-lab.de

ABSTRACT

The number of new mobile and wearable technologies with built-in sensors for quantifying every aspect of our lives is increasing. Consequently, new data sources and opportunities arise for the development of machine learning (ML) models and their applications. In this paper, we report on a four weeks field study with 16 older adults, aged between 66 and 81 years (50% female), who were asked to provide stress-related experience samples in different modalities, including paper-based diaries and data collected with the help of a wearable (i.e., a Microsoft Band 2). We provide insights into participants' stress annotation behavior, report on a detailed analysis of the recorded data and the resulting implications regarding the annotation of stressful situations by older adults, discuss how mobile annotation technology can benefit from the synergies with traditional methods and argue why we believe that appropriate annotation techniques are the basis to benefit individually from future powerful machine learning models.

CCS CONCEPTS

• **Human-centered computing → Field studies**; *Mobile computing*; Mobile devices; • **Applied computing → Annotation**; • **Social and professional topics** → Seniors; • **Hardware** → Signal processing systems;

KEYWORDS

Older adults, annotation, stress diary, field study, wearable, health assistance, experience sampling, signal processing

ACM Reference Format:
Michael Dietz, Ilhan Aslan, Dominik Schiller, Simon Flutura, Anika Steinert, Robert Klebbe, and Elisabeth André. 2019. Stress Annotations from Older Adults - Exploring the Foundations for Mobile ML-Based Health Assistance. In *The 13th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth'19), May 20–23, 2019, Trento, Italy.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3329189.3329197

PervasiveHealth'19, May 20–23, 2019, Trento, Italy

1 INTRODUCTION

Figure 1: Participant using a wristband to create annotations

The term personal informatics (e.g., [19, 20]) was initially used to describe an emerging class of systems, which focused on collecting personal information and consequently improving self-knowledge. About a decade later, the technological landscape seems to have changed and one could argue that personal informatics has become the default kind of everyday informatics for users of all walks of life, young and old. Bulks of personal data are being generated and collected by multiple sources, including social networks and a growing number of mobile and smart technologies augmented with sensing capabilities. Arguably, the health domain is the most promising area for such a form of digitalization, which generates personal data to provide a clearly positive impact on people's lives. But data seems meaningless if it is not properly labeled and contextualized for powerful technologies, such as machine learning techniques. And older adults, who might arguably benefit most from new forms of attentive and "quality of life" improving technologies may still be the most critical towards technology adoption [4].

Despite obvious obstacles including technology usage itself potentially becoming the source for stress [28], the prospect of "game-changing" health services for older adults has motivated us to explore with 16 older adults in a field study user specific behavior and behavioral data, focusing on (i) whether older adults would be willing to use mobile technology to annotate their daily stress experiences, (ii) what patterns would emerge considering older adults' stress experiences within daily and weekly routines, and (iii) how the mobile annotation technology can be complemented with the advantages of traditional approaches such as paper-based diaries. We believe, that it is important to improve our understanding of older adults' annotation behavior and explore new ways for them to provide data and annotations, which are ultimately needed as a basis for new solutions that may contribute to the improvement of their wellbeing, autonomy, and help in dealing with stressful times.

In the field study we applied a mix-measurement method, combining a time-triggered Experience Sampling Method (ESM) [18] on a wristband (i.e., a Microsoft Band 2) with continuous measurements of physiological data from the band and a traditionally paper-based diary method [6]. Our intention was to gather data in comprehensive manner. That is, we combined these methods to make sure that the specificity and quality of collected data was granted. But we also assumed that on the one hand older adults would feel comfortable to provide pen- and paper-based diary entries. On the other hand, through the use of technology, data would be provided in a structured and easy way even if an older participant felt unmotivated to provide detailed diary entries.

Contributions that we hope to achieve through the study and a detailed presentation of the results include insights into the annotation behavior of older adults as well as possible implications for future technology-based solutions which make use of the labeled data to detect stressful experiences and assist older adults in these situations. Before we present the field study and its results in detail we provide in the next section background, including the role of data annotation for active machine learning based health assistance.

2 BACKGROUND

In general, machine learning deals with the problem of designing algorithms that can automatically learn to detect patterns in data and allow a computer to make predictions based on what it has learned. In recent years machine learning methods have been applied successfully to solve various health-related recognition tasks [10, 13, 32] including stress detection [1, 3, 25]. Despite its continuously growing success the method is still subject to some caveats.

Overall, one can distinguish between three broad categories of machine learning algorithms: Unsupervised, Supervised and Reinforced. The main difference between those categories is the way they handle the data to learn from it and the different areas of application that are resulting from that. What all machine learning algorithms have in common is, that the quality of a trained model, with respect to the accuracy of the made predictions at run time, is largely dependent on the data that has been used to train a system. In order to develop a model that performs well during real world usage, it is necessary to train the system on large amounts of prerecorded data that reflect the characteristics of the later analyzed signal as close as possible. However, recording such *in the wild* data is often a highly challenging task in itself.

Another important aspect that can greatly influence the prediction performance of any trained model is the quality of annotations. Depending on the given task and chosen method, the process of annotating data can be rather complex and time consuming. Often large amounts of data are annotated by multiple annotators using specialized annotations tools [9, 17, 31, 34] to obtain the ground-truth. However, this traditional annotation approach depends heavily on the availability of human-comprehensible data (e.g., video or audio) that provides insights into the situative context of the recordings. Furthermore, this approach is limited to the annotation of phenomena that can be observed externally by reviewing the data, which might also lead to annotations that do not correspond with the self-perception of the recorded participants.

Alternatively it is possible to involve the participants directly into the annotation process. A popular approach to do this are diary studies where the participants are asked to report on certain aspects of their daily lives in form of diary entries [7]. A variation of such diary studies is the Experience Sampling Method (ESM) [18]. The peculiarity of the ESM-procedure is that the participants are queried to report on their experiences during the current activity instead of having to reflect on them retrospectively in a diary. In this regard, notifying the participants and reminding them to provide annotations also reduces their burden compared to participants reporting the data on their own accord [8]. Here it is up to the researcher to determine the best moment for notifying the user and collecting the annotations. In general, there are three types of notification strategies to do so: (i) *signal contingent*, in which respondents report when signaled (usually at random times), (ii) *interval contingent*, where annotations are collected at a regular (time-based) interval, and (iii) *event contingent*, in which participants report experience samples in response to certain events of interest [2, 33].

For the present study we selected an interval contingent notification strategy with the option to provide annotations on demand, which allowed participants to also report on rare or irregular events in addition to the regular queries resulting in a comprehensive overview of the experienced situations. Independent of the chosen strategy, the close temporal proximity between an experience that influences the participant's current state of mind and the annotation helps to avoid incorrect situation assessment, caused by erroneous reconstruction of memories [30]. Additionally, the increased annotation frequency allows for a much more fine-grained - and therefore accurate - assessment of the participant's mental state throughout the day.

An annotation concept from the field of machine learning that can be closely connected to ESM is the so called Active Learning (AL) [27]. AL depicts a method, where the machine itself decides which samples are the most effective to learn from and asks an oracle (e.g., the user) to provide a corresponding label for those samples. The goal of this method is to improve the classification accuracy of a model while simultaneously lowering the necessary amount of annotated training data. When active learning is applied to ask the user as soon as new relevant sensor input is detected, it can be viewed as experience sampling with a dynamic querying policy. This state of the art approach to train machine learning models in an online environment has been subject to recent research [23]. For instance, Flutura et al. [14] explored how a generalized base model for detecting drink activity can be personalized and improved by applying this interactive machine learning process. In their study they used a smartwatch to record drinking-related movement data and to collect annotations from the participants which were then used to adapt the model. A similar approach could also be applied to the detection of stress. While previous research already investigated suitable physiological measures for the automated recognition of self-reported stress on mobile devices [15, 26], most of these studies were either conducted with young participants or did not make use of the annotations to personalize the classification models in an online environment. With the present work, we therefore aim to explore the feasibility of collecting stress-related annotations from older adults, which can serve as a foundation for future machine-learning-based systems providing health assistance to older adults in stressful situations.

3 METHOD

3.1 Participants

In order to evaluate the feasibility of the mobile stress annotation approach with older adults, we conducted a field study with 16 participants (50% female) aged between 66 and 81 ($M = 73.3$) years. The criterion for inclusion consisted of a minimum age above 65 years. Candidates with severe affective or cognitive disorders were excluded during recruitment. The sample of participants was well educated (62.5% had a university degree) and rather healthy (56.3% of participants subjectively rated their health condition as "good" or "very good"). Only three participants mentioned an impairment or chronic disease. 50% of the participants were married or had a relationship and none of them was still employed. Regarding their social activities 37.5% specified that they were working as volunteers while 12.5% were members of a club or society. In general, the sample consisted of active, well educated and rather healthy older adults living in an urban environment.

3.2 Procedure

At the beginning of the study we invited each of the 16 participants to an initial examination at our lab. During this visit participants received a brief overview about the details and the procedure of the study and were asked to fill out a questionnaire regarding sociodemographic data. Following this, participants were given a sensor wristband as well as a smartphone and were instructed on how to operate and use them to record stress annotations and physiological data within the scope of the study. This included a detailed briefing of all necessary steps to perform each task, which were also documented in a custom manual that was given to the participants. Additionally, they received a stress diary and instructions on how to fill out the handwritten protocols. Thereby we let the participants decide, whether they wanted to log the stressful situations of their daily lives immediately after they happened or in retrospect every evening. The only requirement was that they should wear the wristband, record annotations and fill out the diary every day for the next four weeks (28 consecutive days). After this period we invited them again for a final examination at our lab. There we asked them several questions regarding the usage of the system and let them fill out a questionnaire about its usability.

3.3 Measures

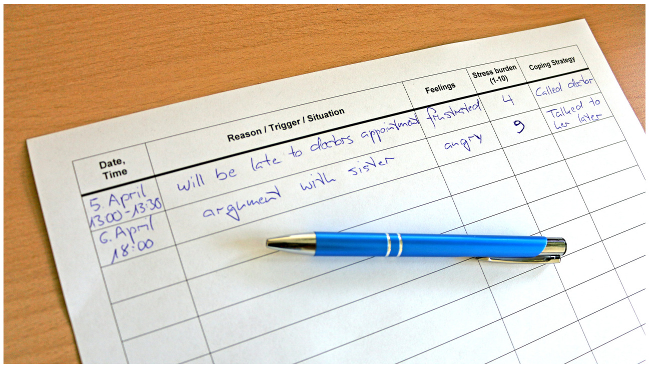
Within the scope of the study several measurement instruments were used. The details of each of them are described in the following paragraphs.

3.3.1 *Demographics Questionnaire.*

The questionnaire to collect sociodemographic data included the following items: *age, gender, martial status, highest educational achievement, number of people living in the same household* and *average net household income*. Additionally, we collected information about the health condition of the participants, such as their subjective assessment of it, whether they suffer from chronic diseases or require a mobility aid. This data was used to create a more differentiated evaluation as well as to investigate correlations between demographics and certain results.

3.3.2 *Stress Diary.*

The stress diary was mainly used as a baseline to identify potential aspects that can be used to complement and improve the mobile stress annotation approach. In addition to that, we also aimed to gain more insights about the circumstances and characteristics of the stressful situations from our target user group. As shown in Figure 2 the diary therefore consisted of the following elements: *date/time, stressor, feelings, stress burden (1-10)* and *coping strategy*. Since the entries from our participants were in free text, we had to categorize them first before we could begin the evaluation. For that, we used a qualitative content analysis [22] with defined encoding rules for each category.



Date, Time	Reason / Trigger / Situation	Feelings	Stress burden (1-10)	Coping Strategy
5. April 13:00-13:30	will be late to doctors appointment	frustrated	4	Called doctor
6. April 18:00	argument with sister	angry	9	Talked to her later

Figure 2: Paper-based stress diary

The *stressors* were divided into the categories *household, health, traffic, technology, interpersonal problems, memory, time pressure* and *physical activity*. While most entries could be distinctly assigned to a specific category, others were more ambiguous. In these cases the *feelings* and *coping strategies* were also considered during the category assignment. For example, the stressor entry "*late for appointment; looking for parking lot*" (P13) could match the categories traffic and time pressure but due to the feeling "*annoyed because of parking situation*" it was assigned to the traffic class. The categorization for the *coping strategies* was adapted from [5] and includes the classes *active solving, active not solving, passive solving* and *passive not solving*.

3.3.3 Sensor Wristband. In order to record mobile stress annotations as well as to measure physiological data, every participant received a Microsoft Band 2 wristband and a Google Nexus 5 smartphone for four weeks. During this time they were told to wear the wristband as often as possible. The reason why we selected the Microsoft Band 2 as opposed to other smartwatches or wristbands is its capability to record the *Galvanic Skin Response (GSR)*, which is rather rare in current devices. Since the battery capacity of the Microsoft Band 2 only lasts about eight hours when constantly recording data, participants were also instructed to charge the devices by themselves, which took approximately one hour before they could use them again. While wearing the wristband, participants received a prompt on its display every hour showing the current time and asking whether they were stressed in the past 60 minutes or not. We chose this duration since it was long enough to not be considered disturbing but also regular enough to still receive meaningful data. The prompt was accompanied by a short vibration to draw the attention of the wearers towards it and could be answered by selecting "yes" or "no" on the touchscreen of the wristband as shown in Figure 3. While this rather simple query does not reveal much about the annotated situation, it was sufficient to evaluate the feasibility of the general mobile annotation approach with older adults.

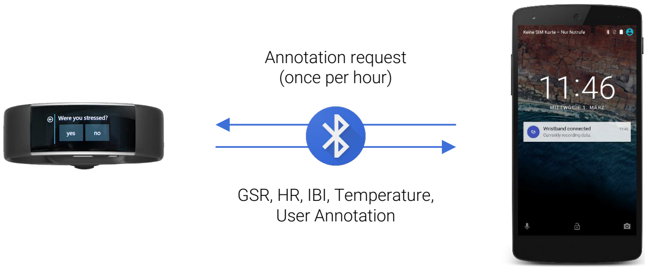


Figure 3: Sensor setup

In addition to the stress annotations, we also recorded the *Galvanic Skin Response (GSR), Heart Rate (HR), Interbeat Interval (IBI)* and *Skin Temperature (ST)* with the integrated sensors of the wristband. For that, we implemented an automated recording system using the SSJ framework [11, 12], which generally enables the recording, processing and classification of social signals on Android smartphones using device internal and external sensors. Due to the modular architecture and component-based nature of the open-source framework we were able to quickly build a signal processing pipeline which recorded the data on the Microsoft Band 2 and transferred it via Bluetooth in real-time to the smartphone where it was stored for later analysis. There, a scheduling component also triggered the hourly stress annotation prompts on the wristband using the same communication channel. Additionally, the system was adjusted to automatically start the recording once the wristband was worn and to stop the data collection as soon as it has been taken off. This prevented notifications at unfavorable times (e.g., during night).

3.3.4 Experience Questionnaire. Within the scope of this questionnaire we collected data about the user experience of the sensor wristband and the annotation system. This included three questions regarding the usability of the overall system, the wristband itself and the vibrations during an annotation prompt. Furthermore, we asked the participants how often they encountered problems while charging the devices, connecting the wristband and the smartphone, and initiating the data recording, which they could answer with *daily, multiple times per week, once a week, less than once a week* and *never.*

4 RESULTS

4.1 Evaluation of Stress Diaries

4.1.1 Stressors. In total participants reported 259 stressful situations in their diaries within the four week period of the study. This corresponds to an average of 16.2 entries per user ($min = 1, max = 45, SD = 12.3$). Based on the responses the most commonly mentioned source of stress was *physical activity* ($M = 3.6$), followed by *technology* ($M = 2.4$), *household* ($M = 2.1$) and *traffic* ($M = 2.1$) as shown in Figure 4. Taking demographic data into account reveals some differences between both genders. While males reported *physical activity* ($M = 5.1$), *technology* ($M = 2.8$) and *traffic* ($M = 2.6$) as most frequent stressors, females mentioned *household* ($M = 3.2$), *health* ($M = 2.3$) and *physical activity* ($M = 2.1$) as their most common sources of stress. However, no significant correlations between the number of entries and demographic characteristics of the participants, such as their age ($r_s(14) = .010, p = .971$), education ($r_s(14) = -.042, p = .878$) or health condition ($r_s(14) = -.182, p = .501$) were found.

Interestingly, participants who assessed their health condition as "very good" ($M = 6.0$) and "rather good" ($M = 4.8$) made more entries for stressful situations caused by *physical activity* than those with an "average" ($M = 1.5$) and "rather bad" ($M = 3.0$) subjective state of health. Additionally, participants with chronic diseases ($n = 3, M = 7.3$) also mentioned *physical activity* as source for stressful situations more often than those without them ($n = 13, M = 2.7$). Overall, the entries in this category covered a wide area of activities ranging from "*wood cutting*" (P8) to "*carried 20kg up 70 stairs*" (P9) and "*drove up a steep hill with my bicycle*" (P5). Regarding the *technology* category, more than half (58.9%) of the stressful situations were caused by the sensor wristband and its empty battery. Other entries included "*problems with printer*" (P1), "*no Skype connection on my PC*" (P2) and "*repair costs for notebook higher than expected*" (P16). In the *household* category most of the entries revolved around grocery shopping, cooking and cleaning. On average, females ($M = 3.2$) reported more than twice as much about stressful situations caused by household problems than males ($M = 1.1$). Similarly, participants living in a two-person household ($n = 7, M = 3.2$) also made more entries in this category than participants living alone ($n = 9, M = 1.3$).

4.1.2 Stress Burden. In addition to reporting the stressors, participants were also able to rate the perceived stress burden of each

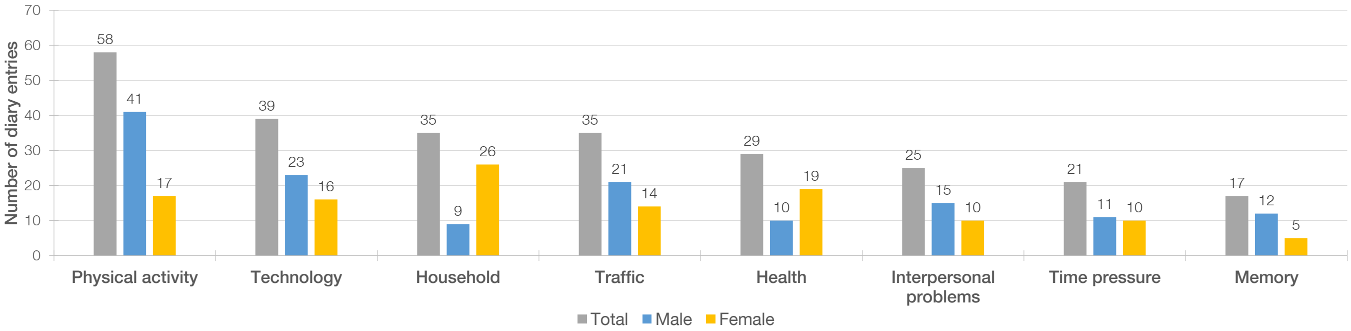


Figure 4: Distribution of diary entries across stressor categories

situation on a scale from 1-10 (with 10 being the highest burden). Though, since not every participant assessed all recorded situations, only 214 out of 259 diary entries were rated ($M = 13.3, SD = 9.7$). On average, the reported stress burden was 4.6 ($min = 0, max = 7.5, SD = 1.9$). In regard to the gender, males ($M = 5.0$) reported a higher average stress burden and also rated more situations ($M_m = 16.2, M_f = 10.5$) compared to females ($M = 4.1$). Considering the level of education, participants with a university degree reported the highest stress burden ($M = 5.0$). Furthermore, participants who were active members of a club or society had a higher stress burden ($n = 2, M = 7.2$) than participants without such activities ($n = 14, M = 4.2$).

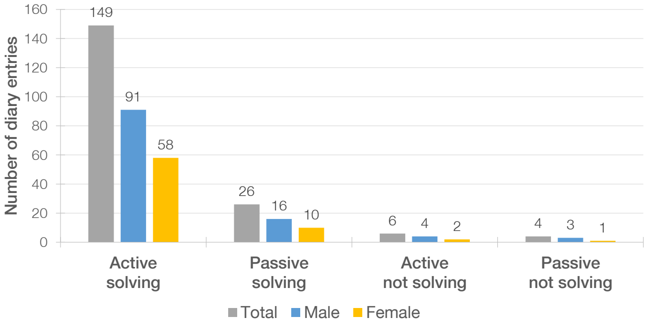


Figure 5: Distribution of diary entries across categories for coping strategies

4.1.3 Coping Strategies. Besides the stress burden, participants could also report how they coped with certain stressful situations. Overall this was done 189 times ($M = 11.8, SD = 8.2$). On average, males reported their coping strategy in 14.7 entries ($SD = 8.1$) while females recorded it in 8.8 cases ($SD = 7.7$). As shown in Figure 5, most of the entries could be classified into the category "*active solving*". The criteria for that was the presence of an active behavior visible to others aimed towards solving the problem which the participant was facing during the stressful situation. Examples for that are: "*Called the doctor*" (P3), "*Looked for alternative traveling options*" (P7) and "*Wrote a message to the neighbors*" (P8). In total 148 entries were grouped into the "*active solving*" category ($M = 9.3, SD = 7.6$). Participants with a high school diploma ($n = 3, M = 15.3$) reported on average the most entries in this category, compared to participants with a university degree ($n = 10, M = 8.8$) and a secondary school certificate ($n = 3, M = 5.0$). Besides actively solving the problems, participants also stayed passive in 26 reported cases to resolve the situations ($M = 1.6, SD = 1.4$). Only in ten cases they did not try to improve their conditions by either doing nothing or doing something which did not help the situation (e.g., complaining about something which can not be changed).

4.2 Evaluation of Wristband Data

4.2.1 Data Overview. Within the scope of the study we collected more than 2484 hours of data per modality (GSR, HR, IBI and ST). The details for each participant are shown in Table 1. On average, each user wore the wristband around 21 days ($min = 4, max = 28, SD = 7.4$) which translates to 7.2 hours per day. During this time they made 5.6 annotations per day, resulting in a total of 1967 collected annotations over the course of the study. Out of those annotations 195 were labeled with "stress" (10%).

4.2.2 Annotation Count Based on Demographics. When taking a closer look at the demographics, it appears that on average males ($M = 118.8$) used the "no stress" label more frequently than females ($M = 102.8$). However, for the "stress" label there is almost no difference between males ($M = 12.1$) and females ($M = 12.2$). Considering the health condition, participants who assessed their state of health as "rather bad" ($n = 1, M = 29.0$) or "average" ($n = 6, M = 13.5$) labeled more annotations with "stress" than users with a "very good" ($n = 2, M = 9.0$) or "rather good" ($n = 7, M = 9.6$) subjective health assessment ($r_s(14) = .235, p = .381$). In return participants with a "very good" ($M = 133.0$) health condition used the "no stress" label the most, compared to those with a "rather good" ($M = 100.0$), "average" ($M = 114.3$) and "rather bad" ($M = 120.0$) state of health.

4.2.3 Physiological Differences Based on Demographics. Participants who assessed their health condition as "very good" ($M = 65.5$) had the lowest heart rate compared to participants with a "rather good" ($M = 72.1$), "average" ($M = 73.2$) and "rather bad" ($M = 75.7$) assessment of their state of health ($r_s(14) = .410, p = .058$). Despite the rather long annotation duration of one hour, there was a significant difference ($t(15) = 2.39, p = .03$) between the heart rates during the "stress" ($M = 73.7, SD = 3.8$) and "no stress" ($M = 72.2, SD = 4.1$) annotations. Additionally, while the heart

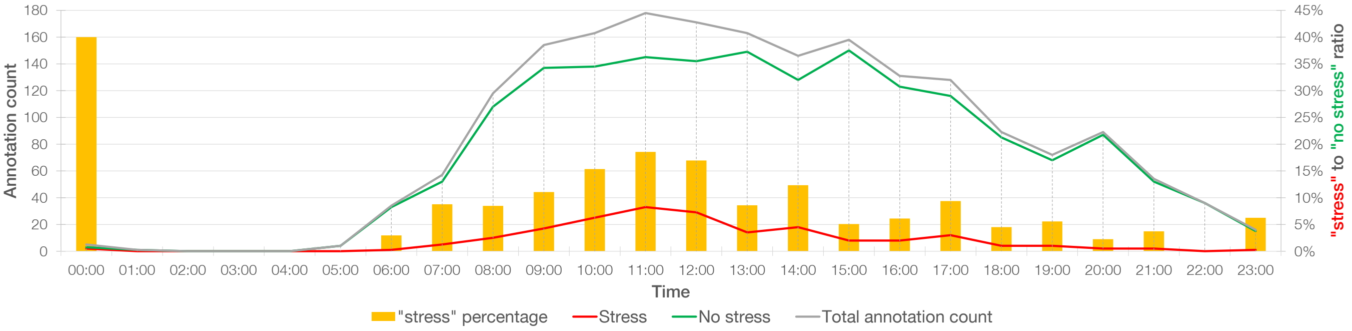


Figure 6: Stress annotation distribution across hours of a day

rate of males ($M = 72.7$) was lower than that of females ($M = 74.4$) during “stress” annotations, it was higher in “no stress” situations ($M_m = 72.6, M_f = 71.9$). Although the other physiological measures also showed some differences between the “stress” and “no stress” annotations, none of them were significant (GSR: $t(15) = -1.081, p = .297$; IBI: $t(15) = -1.660, p = .118$; ST: $t(15) = -.319, p = .754$). Considering demographics, participants who assessed their health condition as “very good” ($M = 1343.4$) had the lowest GSR compared to those with a “rather good” ($M = 5588.4$), “average” ($M = 3250.9$) and “rather bad” ($M = 4805.2$) subjective rating of their state of health.

	Recordings			Annotations			
	Hours	Days	Hours / day	Stress	No stress	Total	Anno / day
1	254.6	27	9.4	5	147	152	5.6
2	31.1	8	3.9	1	1	2	0.2
3	155.3	27	5.7	9	118	127	4.7
4	218.8	26	8.4	6	148	154	5.9
5	111.6	16	6.9	6	78	84	5.2
6	26.6	4	6.6	1	22	23	5.7
7	257.4	27	9.5	26	195	221	8.1
8	133.5	16	8.3	8	97	105	6.5
9	76.1	11	6.9	21	39	60	5.4
10	169.0	22	7.6	13	119	132	6.0
11	184.1	26	7.0	31	135	166	6.3
12	155.3	22	7.0	9	116	125	5.6
13	176.8	26	6.8	29	120	149	5.7
14	98.5	27	3.6	10	207	217	8.0
15	178.3	24	7.4	12	162	174	7.2
16	256.9	28	9.1	8	68	76	2.7
Avg.	155.2	21	7.2	12	110	122	5.6
Sum	2484.6	337	114.8	195	1772	1967	89.5

Table 1: Recorded data from the sensor wristband

4.2.4 Time-Based Annotation Distribution. In addition to the demographic analysis we also evaluated the time-based distribution of annotations. The majority of annotations were made in the daytime between 8 a.m. and 5 p.m. as shown in Figure 6. While the number of “no stress” annotations remains rather constant during noon, the “stress” annotation count steadily rises until it reaches its peak at around 11 a.m. This is also reflected in the percentage ratio between “stress” and “no stress” annotations, which reaches a peak value of 18.5% at this time. With the exception of two smaller peaks at 2 p.m. and 5 p.m., the number of “stress” annotations steadily declines until 10 p.m. after that.

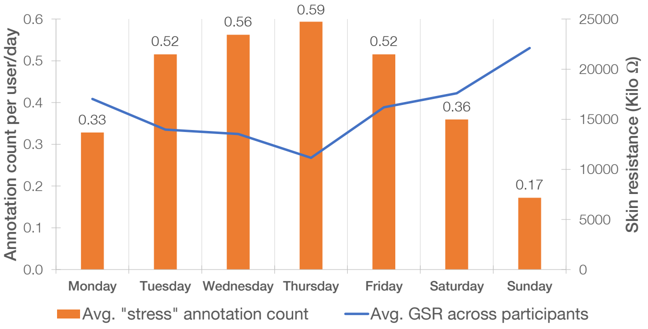


Figure 7: Average “stress” annotation count and GSR distribution per participant across weekdays

When looking at the “stress” annotations on a weekly dimension as shown in Figure 7, we can see that the average annotation count per user is rather low (0.33) on Mondays and increases throughout the week until it reaches a peak of 0.59 on Thursdays. As the weekend begins, the average amount of “stress” annotations per day and user decreases again until it reaches the lowest count of 0.17 on Sundays. This is also reflected in the percentage ratio between “stress” and “no stress” annotations, which starts at 9.2% on Mondays, increases to 12.4% on Thursdays and decreases to 4.7% on Sundays. Although none of the participants were still employed, the recorded distribution looks like one we would expect from regular employees [29]. A similar trend can be observed in the average GSR value across each weekday. Since the Microsoft Band 2 measures the skin resistance, a lower value means that the wearer is sweating more which can be an indicator of a higher stress level. As we can see in Figure 7 the average GSR starts on a medium level on Mondays and steadily decreases until it reaches a minimum value on Thursdays. After that it rises again and reaches a maximum value on Sundays. Results of the Spearman correlation indicate a significant inverse correlation ($r_s(14) = -.955, p = .001$) between the subjective annotations and the objective physiological data (GSR).

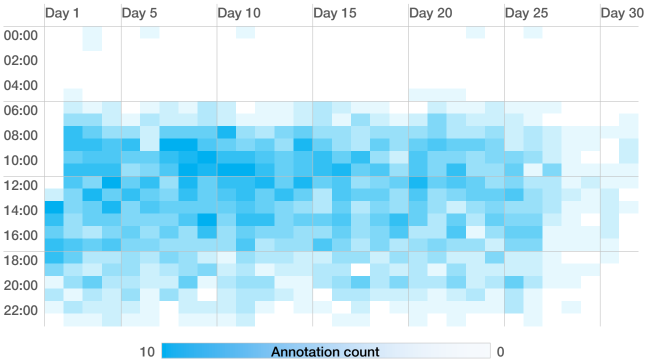


Figure 8: Annotation distribution across study duration

An analysis of the recorded annotation data on a broader scale across the complete duration of the study reveals, that after a certain point the number of annotations decreases over time as shown in Figure 8. While the daily count remains rather constant until day 20 we can observe a moderate decline until a larger drop off occurs around day 25. After that the amount of recorded annotations shrinks drastically until the end of the study. It also appears that the decrease occurs during all hours of the day which results in few annotations during the peak time at 11 a.m. and no annotations during the evening and night.

4.3 Combined Evaluation

4.3.1 Annotations Matching Diary Entries. Based on the individual results of the stress diary and the wristband evaluation, we analyzed the relations and connections between them. One of the most apparent aspects for that was the accordance of stress annotations with diary entries. Since the accuracy of diary entries varied across participants, we grouped them into the following categories of accordance: *entry within ±2h window*, *entry on same day* and *no entry*. As displayed in Figure 9 the date and time of 97 "stress" annotations ($M = 6.1$) matched with a diary entry within a ±2h window. In 49 cases ($M = 3.1$) the annotations corresponded to a situation in the diary reported on the same day while in another 49 cases ($M = 3.1$) no matching entry could be found. Regarding the gender, males ($n = 58, M = 7.3$) reported more situations in their diaries which corresponded to an annotation within a ±2h window than females ($n = 39, M = 4.9$). However, females ($n = 42, M = 5.3$) labeled more "stress" annotations which did not match a diary entry than males ($n = 7, M = 0.9$).

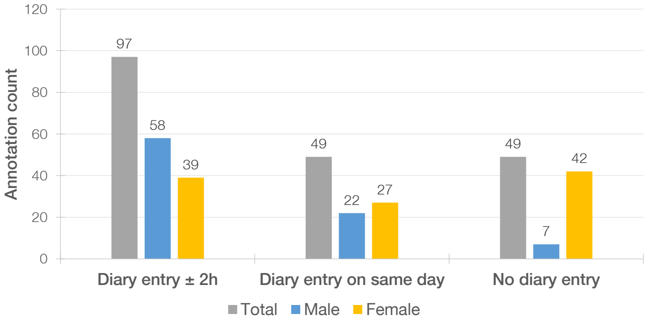


Figure 9: Annotations matching diary entries

4.3.2 Heart Rate Based on Stressors. In order to gain a more detailed overview of the situations with corresponding annotations we then analyzed the sensor data during the 97 entries with a matching annotation within a ±2h window. As shown in Figure 10, participants had the highest average heart rate (76.4 bpm) during health related stressful situations. Following that are circumstances caused by physical activity (74.7 bpm), memory issues (74.6 bpm) and traffic (73.5 bpm). The lowest average heart rate (70.2 bpm) was observed during household related situations.

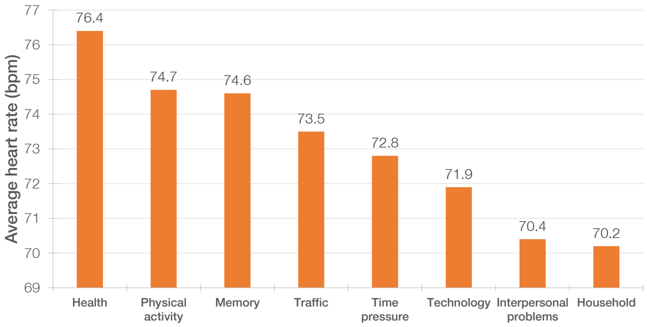


Figure 10: Heart rate based on stressor category

4.3.3 Heart Rate Based on Stress Burden. In terms of stress burden the average rating across the 97 entries was 4.8 out of 10. Since this rating is close to the mean value of the scale, we divided it into the categories *low burden* (1-5) and *high burden* (6-10) for further analysis. As it turns out, most of the entries ($n = 59$) could be assigned to the *low burden* category. Only 35 situations were rated with a stress burden of 6 or higher. Regarding the average heart rate during these categories, we found that in situations with *low burden* it was lower ($M = 71.7, SD = 4.7$) than in situations with *high burden* ($M = 76.0, SD = 5.9$). This also indicates a conformity of the subjective ratings from the participants with the objective measures from the sensors.

4.4 Annotation Experience

In general, the amount of collected data indicates that the mobile annotation approach using a wristband was rather effective. During the usage of the device throughout the study though, some participants experienced issues. 43.7% of the users faced problems at least once a week when charging the wristband. 12.5% even had these problems daily. Additionally, more than half of the participants (68.8%) occasionally (less than once a week) had issues with the data recording. However, despite all of these problems 56.3% of the participants rated the wristband as "very" or "rather" user-friendly. Only 18.8% of the participants found the wristband to be "not very" user-friendly. Regarding the gender, males (62.5%) rated the device more frequently as "very" or "rather" convenient than females (50%). Similarly, more than half of the participants (62.5%) found the hourly stress annotation request to be "very" or "rather" user-friendly. In this regard almost all participants (93.8%) rated the accompanying vibration as appropriate and did not assess it as annoying or disturbing.

5 DISCUSSION

Having provided a detailed evaluation we discuss in this section the results, including their implications for future research and the developments of mobile health assistance for older adults.

5.1 Potential Implications for Machine Learning Applications

The collected amount of data indicates a rather high level of compliance among the participants of our study to wear a wristband regularly and record annotations with it. While these preliminary results should be confirmed on a larger scale, they suggest that older adults seem willing to use mobile technology to annotate their daily stress experiences. Future machine learning based assistance systems could take advantage of this finding by enabling the users to personalize the models, responsible for detecting critical situations, with their own annotations. This could potentially improve the recognition accuracy and could increase the users' trust towards such systems.

In addition to that, we observed a relation between certain subjective assessments and objective physiological measures which could further contribute to the improvement of recognition models. More precisely, we found a correlation with medium effect size between the subjectively rated health condition and the average heart rate of the participants ($r_s(14) = .410, p = .058$). This indicates that incorporating the health condition as criteria into the machine learning process could improve the results. For instance, one possible approach would be to use multiple models (one for each health condition). The reason for this is the presence of large differences between the average heart rates in relation to the subjective health assessments of the participants. Therefore, when only using one model the selected features have to account for these variances, which often leads to lower recognition rates. When using multiple models though, the differences within one condition are much smaller which facilitates the detection of anomalies and might lead to better recognition rates. However, since the correlation between the health condition and the average heart rate was not significant and only based on a small sample size, these hypotheses have to be verified in larger scale studies.

5.2 Insights on Heart Rate and Stress

One measure where we found a significant difference was the average heart rate during "stress" and "no stress" annotations ($t(15) = 2.39, p = .03$). This result indicates that the heart rate signal could be used to distinguish both classes with machine learning methods. Although the absolute difference seems rather small, its presence despite the relatively long annotation duration of one hour has to be considered. For smaller annotation windows we would expect to find even larger differences. Therefore, creating features based on the heart rate data might yield a solid foundation for future classification models aimed towards detecting stress in our target user group. Additionally, we found that the average heart rate of males was lower than that of females during stressful situations, which corresponds to the results of previous studies [21]. While the other physiological measures also showed some differences between both labels, none of them were significant, which is not unusual considering the rather long annotation duration of one hour. In future studies it would be advised to investigate shorter annotation durations as well, which might lead to further results considering the other physiological measures as indicated by fellow researchers [15, 26].

5.3 Times of "Habitual" Stress

5.3.1 "Habitual" Stress Across a Week. One interesting observation we made regarding the weekly distribution of "stress" annotations was, that it starts on a rather low count on Mondays, then rises until its peak on Thursdays and afterwards decreases until it reaches the lowest count on Sundays (Figure 7). While such a distribution might be expected from students and employees as shown in [29], this explanation can not be used in our present study since the participants were all retired. One possible explanation could be that the participants followed a certain routine during their working life and continued it even after their retirement. Another explanation could be that despite being retired, participants still got in contact with the working population (e.g., shopping, doctor's appointment, etc.) which caused a similar amount of stressful situations. This distribution is also reflected in the physiological data where we found an inverse correlation between the average GSR and the "stress" annotation count across each weekday ($r_s(14) = -.955, p = .001$). The correlation of physiological data with subjective annotations indicates that the GSR might be a useful measure to predict the number of stressful situations on a given weekday. It also indicates that the participants had a rather accurate self-assessment ability which contributes to the effectiveness of the mobile annotation approach aimed towards creating personalized machine learning models for older adults.

5.3.2 "Habitual" Stress Across a Day. Considering the daily annotation distribution, we observed a similar progression as in the weekly one with a majority of annotations between 8 a.m. - 5 p.m. and a peak amount of "stress" annotations at around 11 a.m. (Figure 6). This is different from previous studies with young adults [29] where the highest number of labels was recorded between 3 p.m. and 5 p.m. while the majority of annotations were reported between noon and 8 p.m. One reason for these differences could be the distinct daily rhythms between both user groups which might be influenced by the age and employment status of the participants. Therefore, this result also indicates the importance of recording data from the intended users in order to collect the specific characteristics of this group such as with the daily annotation distribution in this case. Combined with the weekly distribution, a reference model containing the probability for the occurrence of a stressful situation at a given time and weekday could be created. This model could then be used in situations where a future machine learning system has a low confidence in its current prediction about the presence of stress. For instance, if the system is unsure about its prediction on a Thursday at 11 a.m., then the time-based model would suggest that the presence of a stressful situation is rather likely. This could either directly influence the prediction of the system or could trigger an annotation request for the user to provide the appropriate label for the given situation.

In this regard it is important that the system adapts to the user input so that a noticeable relation between the provided annotations and the model performance can be observed. Otherwise, the users might lose interest in answering future annotation requests if it yields no benefits for them. An indication for this behavior can be found in the present study, where we observed a steady decline of recorded annotations after a certain point in time (around day 20). Since the system only recorded data, there was no further incentive for the participants to continue the labeling process, which should be considered in future studies.

5.4 Lessons Learned from Combing the Diary and Experience Sampling Method

In general, the evaluation of the paper-based diary entries revealed some valuable insights about the stressful situations, which our participants were facing over the course of the study. Therefore, instead of just asking whether they were stressed or not it might be beneficial to also give the users the ability to categorize the current situation and to rate the stress burden on the wristband. For instance, Hernandez et al. [16] used a 5-point Likert scale ("*How stressed are you feeling right now?*") and a 2D grid with valence and arousal as dimensions for the annotation of stressful situations on different mobile devices. While a Likert scale might be acceptable for our target user group, a 2D grid could already be too complex and discouraging for older adults in the given context. Thus, a balance between complexity and information diversity has to be found.

In addition to the stressors, the diary entries also revealed valuable information about the coping strategies of the participants in the reported situations. As it turns out, a majority of participants already actively performed certain actions to resolve the problems causing stressful circumstances. However, in some cases, the issues were not resolved which resulted in continued stressful experiences. This is where a personalized machine learning based system could automatically detect the problematic situations and provide the users with assistance. In order to reach this goal, the number of collected "stress" annotations needs to be improved first. While it should be easier to label an annotation on the wristband than to fill out the diary by hand, we collected more diary entries (259) than "stress" annotations (195). Additionally, 49 annotations did not match a diary entry while for 113 reported situations no matching annotation could be found. A possible explanation for this could be that the participants were too engaged in the stressful situations and did not think about creating an annotation during the heat of the moment. In contrast, most participants made the diary entries retrospectively which allowed them to reflect on more situations. Therefore, the annotation approach should be extended with an option to also create annotations retrospectively. This change could contribute to an overall higher number of annotations with more accurate time frames as shown in [29] where the majority of annotations were labeled retrospectively.

5.5 Potential Limitations

The present study has provided various insights about the mobile stress annotation behavior of older adults and how it can be complemented with techniques from diary-based approaches. While the number of participants was rather low compared to previous works [30], most of these studies were conducted with students and research assistants who are much easier to recruit than older adults. Despite the small number of participants, the rate of more than 7.2 hours of data recorded per user, modality and day was higher than in most previous studies [15, 24, 26, 29].

Another interesting finding is that more than half (58.9%) of the stressful situations reported in the *technology* category were caused by the empty battery of the sensor wristband. Although this only corresponds to 8% of the total entries, it is still alarming that the system, aimed towards preventing stress in the future, was causing it in these situations. In order to extend the battery life of the Microsoft Band and to prevent further stressful experiences, the data recording of the skin temperature and the IBI could be disabled, since they did not provide much additional insight about the distinctive characteristics of stress in our present study. Alternatively, another sensor wristband with higher battery capacity should be selected since previous works considered the skin temperature and IBI as valuable measures in their stress detection approaches [15, 26].

6 CONCLUSION

While the domain of health may only be one of many areas for future machine learning and data annotation based software solutions, both the increasing amounts of personal data that is being collected in various forms and the importance of future digital health services for an aging society has motivated the work at hand. We have argued how personal data is taking a central stage position in the practice of programming with a shift towards machine learning based software solutions. We believe that these solutions would benefit from HCI research taking a turn towards data practices of special user groups including contextual annotation practices.

To this end, we reported on a field study with older adults, exploring various aspects of their annotation behavior and stress related data. As indicated by the study results, older adults seem willing to use mobile technology to provide annotations. Moreover, the combination of the mobile annotation technology with a traditionally paper-based sampling method revealed valuable insights regarding the experienced situations of older adults and the resulting requirements of annotating them. The detailed analysis and interpretation of the recorded data could also be useful for design and timing choices in future machine learning based solutions to detect stressful experiences and provide assistance for older adults. We sincerely hope that our findings will inspire fellow researchers and practitioners, and ultimately contribute to improving the quality of life for older adults including our future selves.

ACKNOWLEDGMENTS

This work was partially funded by the German Federal Ministry of Education and Research (BMBF) under project EmmA (FKZ 16SV8028) and Glassistant (FKZ 16SV7270).

REFERENCES

[1] Ane Alberdi, Asier Aztiria, and Adrian Basarab. 2016. Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. *Journal of biomedical informatics* 59 (2016), 49–75.

[2] Lisa Feldman Barrett and Daniel J. Barrett. 2001. An Introduction to Computerized Experience Sampling in Psychology. *Social Science Computer Review* 19, 2 (2001), 175–185. https://doi.org/10.1177/089443930101900204

[3] Shaibal Barua, Shahina Begum, and Mobyen Uddin Ahmed. 2015. Supervised machine learning algorithms to diagnose stress for vehicle drivers based on physiological sensor signals.. In *pHealth*. 241–248.

[4] Anabela Berenguer, Jorge Goncalves, Simo Hosio, Denzil Ferreira, Theodoros Anagnostopoulos, and Vassilis Kostakos. 2017. Are Smartphones Ubiquitous?: An in-depth survey of smartphone adoption by seniors. *IEEE Consumer Electronics Magazine* 6, 1 (2017), 104–110.

[5] Kira S Birditt, Karen L Fingerman, and David M Almeida. 2005. Age differences in exposure and reactions to interpersonal tensions: a daily diary study. *Psychology and aging* 20, 2 (2005), 330.

[6] Niall Bolger, Angelina Davis, and Eshkol Rafaeli. 2003. Diary methods: Capturing life as it is lived. *Annual review of psychology* 54, 1 (2003), 579–616.

[7] Scott Carter and Jennifer Mankoff. 2005. When participants do the capturing: the role of media in diary studies. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 899–908.

[8] Yung-Ju Chang, Gaurav Paruthi, and Mark W. Newman. 2015. A Field Study Comparing Approaches to Collecting Annotated Activity Data in Real-world Settings. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 671–682. https://doi.org/10.1145/2750858.2807524

[9] Roddy Cowie, Martin Sawey, Cian Doherty, Javier Jaimovich, Cavan Fyans, and Paul Stapleton. 2013. Gtrace: General trace program compatible with emotionml. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 709–710.

[10] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication* 71 (2015), 10–49.

[11] Ionut Damian, Michael Dietz, and Elisabeth André. 2018. The SSJ Framework: Augmenting Social Interactions Using Mobile Signal Processing and Live Feedback. *Frontiers in ICT* 5 (2018), 13. https://doi.org/10.3389/fict.2018.00013

[12] Ionut Damian, Michael Dietz, Frank Gaibler, and Elisabeth André. 2016. Social Signal Processing for Dummies. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016)*. ACM, New York, NY, USA, 394–395. https://doi.org/10.1145/2993148.2998527

[13] Michael Dietz, Daniel Schork, Ionut Damian, Anika Steinert, Marten Haesner, and Elisabeth André. 2017. Automatic Detection of Visual Search for the Elderly using Eye and Head Tracking Data. *KI - Künstliche Intelligenz* 31, 4 (01 Nov 2017), 339–348. https://doi.org/10.1007/s13218-017-0502-z

[14] Simon Flutura, Andreas Seiderer, Ilhan Aslan, Chi-Tai Dang, Raphael Schwarz, Dominik Schiller, and Elisabeth André. 2018. DrinkWatch: A Mobile Wellbeing Application Based on Interactive and Cooperative Machine Learning. In *Proceedings of the 2018 International Conference on Digital Health*. ACM, 65–74.

[15] A. Ghosh, M. Danieli, and G. Riccardi. 2015. Annotation and prediction of stress and workload from physiological and inertial signals. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 1621–1624. https://doi.org/10.1109/EMBC.2015.7318685

[16] 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 (MobileHCI '16)*. ACM, New York, NY, USA, 195–205. https://doi.org/10.1145/2935334.2935340

[17] Michael Kipp. 2014. Anvil: The video annotation research tool. *Handbook of corpus phonology* (2014), 420–436.

[18] Reed Larson and Mihaly Csikszentmihalyi. 1983. The experience sampling method. *New directions for methodology of social & behavioral science* (1983).

[19] Ian Li, Anind Dey, Jodi Forlizzi, Kristina Höök, and Yevgeniy Medynskiy. 2011. Personal Informatics and HCI: Design, Theory, and Social Implications. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11)*. ACM, New York, NY, USA, 2417–2420. https://doi.org/10.1145/1979742.1979573

[20] Ian Li, Anind K. Dey, and Jodi Forlizzi. 2011. Understanding My Data, Myself: Supporting Self-reflection with Ubicomp Technologies. In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp '11)*. ACM, New York, NY, USA, 405–414. https://doi.org/10.1145/2030112.2030166

[21] Mohamed Faisal Lutfi and Mohamed Yosif Sukkar. 2011. The effect of gender on heart rate variability in asthmatic and normal healthy adults. *International journal of health sciences* 5, 2 (2011), 146.

[22] Philipp Mayring. 2010. Qualitative Inhaltsanalyse. In *Handbuch qualitative Forschung in der Psychologie*. Springer, 601–613.

[23] Tudor Miu, Paolo Missier, and Thomas Plötz. 2015. Bootstrapping personalised human activity recognition models using online active learning. In *Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on*. IEEE, 1138–1147.

[24] Rosalind W Picard and Karen K Liu. 2007. Relative subjective count and assessment of interruptive technologies applied to mobile monitoring of stress. *International Journal of Human-Computer Studies* 65, 4 (2007), 361–375.

[25] Akane Sano and Rosalind W Picard. 2013. Stress recognition using wearable sensors and mobile phones. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 671–676.

[26] Akane Sano, Sara Taylor, Andrew W McHill, Andrew JK Phillips, Laura K Barger, Elizabeth Klerman, and Rosalind Picard. 2018. Identifying Objective Physiological Markers and Modifiable Behaviors for Self-Reported Stress and Mental Health Status Using Wearable Sensors and Mobile Phones: Observational Study. *J Med Internet Res* 20, 6 (08 Jun 2018), e210. https://doi.org/10.2196/jmir.9410

[27] Burr Settles. 2012. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6, 1 (2012), 1–114.

[28] Monideepa Tarafdar, Cary L Cooper, and Jean-François Stich. 2017. The technostress trifecta-techno eustress, techno distress and design: Theoretical directions and an agenda for research. *Information Systems Journal* (2017).

[29] Yonatan Vaizman, Katherine Ellis, Gert Lanckriet, and Nadir Weibel. 2018. ExtraSensory App: Data Collection In-the-Wild with Rich User Interface to Self-Report Behavior. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 554, 12 pages. https://doi.org/10.1145/3173574.3174128

[30] Niels van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The Experience Sampling Method on Mobile Devices. *ACM Comput. Surv.* 50, 6, Article 93 (Dec. 2017), 40 pages. https://doi.org/10.1145/3123988

[31] Johannes Wagner, Tobias Baur, Dominik Schiller, Yue Zhang, Björn Schuller, Michel Valstar, and Elisabeth Andre. 2018. Show Me What You've Learned: Applying Cooperative Machine Learning for the Semi-Automated Annotation of Social Signals.

[32] Johannes Wagner, Thiago Fraga-Silva, Yvan Josse, Dominik Schiller, Andreas Seiderer, and Elisabeth André. 2017. Infected Phonemes: How a Cold Impairs Speech on a Phonetic Level. *Proc. Interspeech 2017* (2017), 3457–3461.

[33] Ladd Wheeler and Harry T Reis. 1991. Self-recording of everyday life events: Origins, types, and uses. *Journal of personality* 59, 3 (1991), 339–354.

[34] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*. 1556–1559.