The Supervised Learning Dilemma: Lessons Learned from a Study in-the-Wild

Kristina Kirsten[1(✉)], Robin Burchard[2], Olesya Bauer[1], Marcel Miché[3], Philipp Scholl[4], Karina Wahl[3], Roselind Lieb[3], Kristof Van Laerhoven[2], and Bert Arnrich[1]

[1] Digital Health - Connected Healthcare, Hasso Plattner Institute, University of Potsdam, Potsdam, Germany
kristina.kirsten@hpi.de
[2] Ubiquitous Computing, University of Siegen, Siegen, Germany
[3] Department of Psychology, University of Basel, Basel, Switzerland
[4] ABB Corporate Research, Baden, Switzerland

Abstract. The increasing popularity of conducting studies in real-life settings, known as "studies in-the-wild", is a valuable addition to the traditional controlled clinical trials. These studies enable the observation of long-term effects and account for the complex influences of everyday life. Body-worn sensors facilitate the continuous and unobtrusive collection of motion data in its user's natural, everyday life environment. However, studies in-the-wild require careful planning regarding equipment usability, accessibility, and the creation of efficient study protocols to maximize the quality and output of the collected data. This paper presents insights from our recent study on compulsive handwashing, highlighting the challenges and strategies in study design, implementation, and label acquisition in order to perform supervised machine learning. We present approaches as well as the benefits and limitations of annotating data retrospectively so that participants are impacted minimally during the study. Finally, we list our learning and insights for upcoming studies of that kind.

Keywords: human-activity-recognition · sensor data · wearables · real-life study

1 Introduction

Conducting studies outside the controlled setting, also referred to as "in-the-wild", has gained enormous popularity in recent years. The concept of observational studies in natural environments without intervention began to emerge along with the questions of what studies outside the laboratory should look like [6]. This specific kind of study offers many benefits. On the one hand, findings from clinical trials can be reassessed in everyday life and observed over a longer period to study long-term effects. On the other hand, for many research

O. Konak et al. (Eds.): iWOAR 2024, LNCS 15357, pp. 181–195, 2025.
https://doi.org/10.1007/978-3-031-80856-2_12

questions, the influences of ordinary life play a major role in the results. These influencing factors can only be observed in a real and realistic environment where the participant is not controlled or monitored. Studies that additionally use sensors can also be carried out outside the lab through the rapid development of portable sensors, known as wearables to collect physiological signals and motion data. Wearables, such as smartwatches, make collecting data constantly but unobtrusively in everyday life possible.

Although studies in-the-wild have great advantages, they also require special precautions in their design and implementation. These are characterized by the choice of equipment concerning usability and accessibility for the participant and a study protocol that is simple to implement but provides the desired output.

Wearables can be used to record physiological parameters as well as motion data continuously. The latter can be used to recognize movement patterns and assign activities. The research field of human activity recognition in-the-wild is popular but poses many challenges. Distinguishing activities in everyday life without further (e.g. contextual) information is highly complex. Thus, when it comes to developing a (machine learning) model that recognizes certain activities, researchers still rely on supervised machine learning methods. For those approaches, the beginning and end of an activity needs to be known. These so-called labels can be received from the study participants during data collection. Nevertheless, it is a significant effort for the participant to provide information not only about the time of an activity but also about the duration, i.e. start and end. However, since we want to influence the subject's everyday behavior as little as possible, we often accept collecting only the information about the time of the activities. This means that for the supervised machine learning model, it is necessary to find a way to determine the start and end time of the activity after the data has been recorded.

In this paper, we show what we have learned from our recent study on (compulsive) handwashing in terms of study design, implementation, and label acquisition. We explore various strategies for addressing noisy labels, coming from real-life situations, in supervised machine learning. We delve into the insights and impacts of manual inspections and annotations and discuss the Inter-Annotator Agreement (IAA). Lastly, we share our lessons learned and insights to guide future research in this area.

2 Background

Since the present work is essentially concerned with the aspects of studies outside the laboratory and the associated challenge of data annotation, we will explain the fundamental aspects below.

2.1 Studies in-the-Wild

Over a decade ago, researchers identified the need for "in-the-wild" studies with wearable devices, arguing that research in participants' natural environments

is crucial for understanding the real-life impact of technology and minimizing behavior changes from observation awareness [6,9].

Schlögl et al. emphasized the importance of involving more users in these studies to validate data from wearable technologies. Their research highlighted that real-life interactions with wearables are affected by technical knowledge and device discomfort, recommending a user-centered approach [15].

Overall, scientific literature indicates that including participants as early as possible in the study design is essential for understanding their needs, ensuring compliance, and achieving high-quality data in unsupervised, real-life studies.

2.2 Time-Series Data Annotation and IAA

Even though research in the area of machine learning is increasingly moving in the direction of deep learning, there are still use cases that are better suited to classic machine learning due to their novelty and limited amount of data. When talking about classic machine learning, a distinction is made between supervised, semi-supervised, and unsupervised learning. While the former requires a considerable amount of annotated data, the advantages in terms of accuracy, interpretability, performance evaluation, and reliability make it the preferred choice for many applications, especially those where precision and reliability are of major importance [2,5].

Despite the benefits, obtaining high-quality annotated data in a real-world study with wearables can be challenging when it comes to capturing participants' natural behavior without additional burden. To keep the effort and influence to a minimum, it is common practice to have third parties enrich the data with additional information (annotations, labels) afterwards. In the following, the terms annotation and labeling are used synonymously.

Annotating data retrospectively by external persons, e.g. by visual inspection, involves a certain risk, especially concerning the introduction of a personal bias. Several annotators are often used for the same data to keep this to a minimum. Approaches such as the calculation of IAA are used to evaluate the success and degree of agreement of the manually annotated data. Prominent metrics for calculating the IAA are, i.e., the Percent Agreement, Krippendorff's Alpha, and Cohen's Kappa (κ) [7]. The latter will be used in this paper and is defined as follows:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o is the observed agreement and P_e is the expected agreement by chance.

3 Related Work

In their systematic literature review, the authors of [1] explicitly highlight that one of the primary obstacles in real-world activity recognition is the necessity for labeled data. However, the review did not identify a simple, high-quality label creation solution. As a possible solution, the paper by Garcia-Ceja and Brena

is referenced, where the authors recommend labeling only a small part of the dataset with available annotated data and using it to train personalized models, as these outperformed general models [8].

The paper by Larradet et al. explores various aspects, including the challenges associated with self-reporting for emotion recognition in daily life. They draw attention to the subjective nature of labeling, which is influenced by individual perceptions. Moreover, they point out the likelihood of delays or inaccuracies in annotations due to the dynamic and unpredictable nature of everyday activities. To address these challenges, they propose implementing standardized annotation protocols to improve consistency and objectivity [13].

In a recent study presented in [11], conducted in real-world settings, the authors highlight significant challenges arising from delays encountered at various stages of the project. These delays range from acquiring ethics approval to facing technical difficulties upon the initiation of the main study. They suggest conducting pilot studies with distinct goals, e.g., to validate assumptions made in the study design or to test the devices in action. This again shows the need to involve future study participants in the planning phase, which is known as a user-centered approach.

4 Previous Works

In a series of previous studies [3,18,19] in preparation for the study presented in this paper, lab-recorded inertial measurement unit (IMU) data from the wrist, collected under controlled conditions, was used to simulate specific handwashing behaviors as it occurs in people suffering from obsessive-compulsive disorder (OCD). The researchers used detailed scripts of compulsive handwashing, based on descriptions from individuals with OCD, to enact specific sequences of handwashing gestures. The goal was to demonstrate that simulated compulsive handwashing could be distinguished from routine handwashing in healthy participants. This approach could later be utilized to support and enhance conventional therapies, such as psychotherapy or specifically exposure and response prevention (ERP) therapy, by automatically detecting and logging compulsive actions, and providing feedback to the patient to help them discontinue these behaviors.

In a subsequent study, the dataset was expanded to include other repetitive activities similar to handwashing [16]. This enhancement aimed to improve the robustness of the trained models against confounding activities, such as “rinsing a cup” or “peeling a carrot” which involved repetitive wrist motions resembling handwashing. The researchers successfully showed that simulated compulsive handwashing could be distinguished from these confounding activities, including routine handwashing.

These pilot studies in the laboratory served both technically and in terms of study design as the basis for the study-in-the-wild presented in the following.

5 Dataset Generation

The collected data is part of a study called OCDetect about compulsive handwashing conducted in Switzerland. The study was approved by the Ethics Committee of North/West Switzerland (application number 2021-01317). We recruited 30 participants with excessive urge to wash their hands. All participants were examined by trained psychologists and had to complete interviews to guarantee their suitability for the study. To be part of the study, participants had to be aged between 18 and 75, non-suicidal, and meet the criteria for compulsive handwashing. In the end, 22 participants completed the study, and eight dropped out for personal or technical reasons.

The participants were asked to wear a smartwatch for at least six hours a day over a 28-day recording period and follow their normal daily routines. The Android-based smartwatches were adapted with a pre-trained machine learning model to recognize handwashing in daily life automatically. Therefore, we recorded the data from the three-axis accelerometer and gyroscope at a frequency of 50 Hz. The model was trained on simulated lab data so that its performance in real life was less than expected. Whenever the watch detected a possible handwashing activity, the user got a notification that could be affirmed or declined. Additionally, the participants could mark handwashing manually by tapping on the watch and also indicate the type of washing, i.e. compulsive or routine handwashing. By this, we received information, later called labels, about the point when a handwashing activity occurred.

We ended up with a cleaned dataset of 2600 h of daily-life activities and a total of 2930 handwashing sessions, of which 1526 were categorized as compulsive by users, while 1404 were identified as routine handwashing sessions.

6 Re-labeling Approaches

For the OCDetect study, we decided to collect only one label in the form of a timestamp for each handwashing event. We aimed to prevent patients with compulsive washing behavior from additional stress and we did not want to change their natural movement patterns unnecessarily. Thereby, we could collect data in a realistic scenario. Consequently, this also means that we have no information afterward about the start and possibly the end of the activity. However, this information is necessary for supervised machine learning approaches. For this reason, we enriched the data with this information retrospectively after completing the data acquisition. Since this step is not trivial without the information about the length of the activity as well as the start and exact end, we considered two different approaches. In the following, these two approaches are referred to as automatic re-labeling and manual re-labeling.

6.1 Automatic Re-labeling

First, we wanted to get a sense of how long participants spend washing their hands on average. Although we found evidence in the literature on the average

duration of hand washing in the German population (more than half of the participants wash their hands for between 10s and 19s on average [14]), we could not assume that this behavior is the same in patients with compulsive handwashing. Therefore, we used the video footage we created during the first visit to the lab, where participants were asked to wash their hands while being filmed. Using the video material, we were able to derive a personal handwashing duration for each participant for whom we had a recording. Unfortunately, this was not the case for every participant, so for those where we had no lab video, we used the average duration of handwashing that we had calculated from all available videos. On average, handwashing in the lab took 38s, with 18s being the shortest and 60s the longest. This observation does not align with the typical handwashing duration seen in the general population, but it supports the naive assumption that individuals with a handwashing compulsion tend to wash their hands for a longer period. Finally, we labeled the activity up to 5s before the actual user label, as we assumed that the hand washing was already over by the time the user pressed the button on the watch to indicate a washing activity had happened.

6.2 Manual Re-labeling

As an additional approach, we opted for an elaborate manual approach to evaluate the extent to which the human factor can improve the labels and thus the result. Since manual annotation is very time-consuming, we decided to annotate only a subset of six participants manually, but with higher quality and less potential bias, rather than relying on quantity. This subgroup already has a sufficient number of participants and annotations to get a feel for the impact on the classification results (which will be published elsewhere).

To minimize personal bias, we opted to use two different annotators for each participant to manually label the handwashing events. With six participants and four annotators, we established unique participant-annotators pairs, we formed the following set of possible labeling assignments:

$$\mathcal{A} = \{(P_i, (A_j, A_k))|i = 1, 2, ...6, j, k \in \{1, 2, 3, 4\}, j \neq k\} \tag{1}$$

We then selected assignments from $\mathcal{A}$ so that the following constraints were met:

$$\forall P_i \in P : P_i \text{ appears in } \mathcal{A} \text{ exactly twice} \tag{2}$$

$$\forall \text{Pairs } (A_j, A_k), j < k : (A_j, A_k) \text{ appears in } \mathcal{A} \text{ exactly once} \tag{3}$$

As an annotation tool, we decided on an open-source, easy-to-use, and collaborative online platform called Label Studio [17]. Since not every handwashing activity is clearly visible, the annotator could choose between four different label types: *Begin AND End uncertain* (if both the activity start and end are difficult to identify), *Begin uncertain* (if only the end can be clearly determined), *End uncertain* (if only the beginning is identifiable), or *Certain* (if the activity is fully recognizable. Additionally, the annotator may opted not to set a manual

Table 1. This table shows the amount of originally set user labels (*before, abbr. as bef.*) and those set by the individual annotator (abbr. as Annot.) for their assigned subject (A - F) afterward (*after*). Each absolute number of before and after labels, as well as the corresponding percentage share, is also provided.

Annot.	**Subject A**			**Subject B**			**Subject C**			**Subject D**			**Subject E**			**Subject F**		
	bef.	*after*	%	*bef.*	*after*	%	*bef.*	*after*	%	*bef.*	*after*	%	*bef.*	*after*	%	*bef.*	*after*	%
1	362	235	65	225	208	92				398	308	77						
2				225	212	94	130	115	89							195	169	87
3							130	127	98	398	347	87	366	343	94			
4	362	195	54										366	323	88	195	124	64

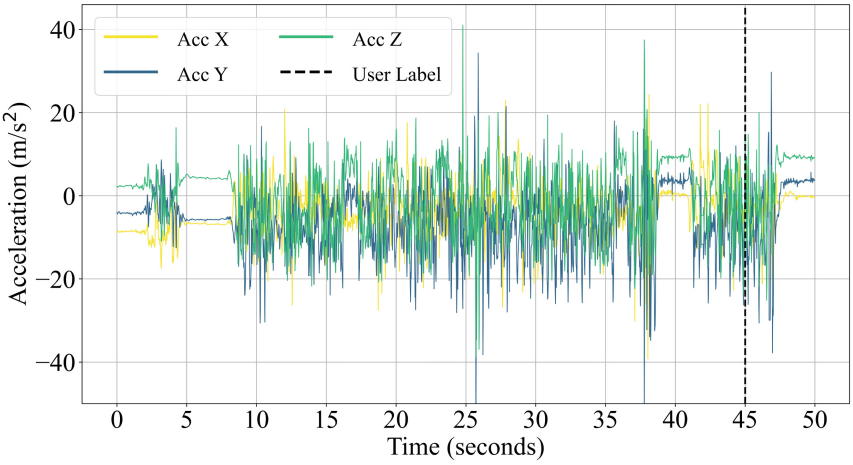
label at all, such as when there is no movement. This differentiation between label types allows for subsequent analyses of the relabeling process.

In Fig. 1, we illustrate visualizations of accelerometer data capturing two distinct handwashing activities performed by Subject E. While in Fig. 1a, the rapid handwashing movement is clearly identifiable, in Fig. 1b, this characteristic is not noticeable. Furthermore, in Fig. 1a, the original user label (depicted by a dotted black vertical line) coincides with ongoing motion, making it challenging to determine whether the activity had already concluded before the movement, merely indicating button pressing, or if the movement still constitutes part of the handwashing activity. However, segments with clearly no movements make it easier to isolate specific patterns, such as handwashing. In contrast, Fig. 1b presents a significant challenge because neither the beginning nor the end of the activity can be distinctly identified visually. This highlights some of the difficulties encountered during manual data annotation.

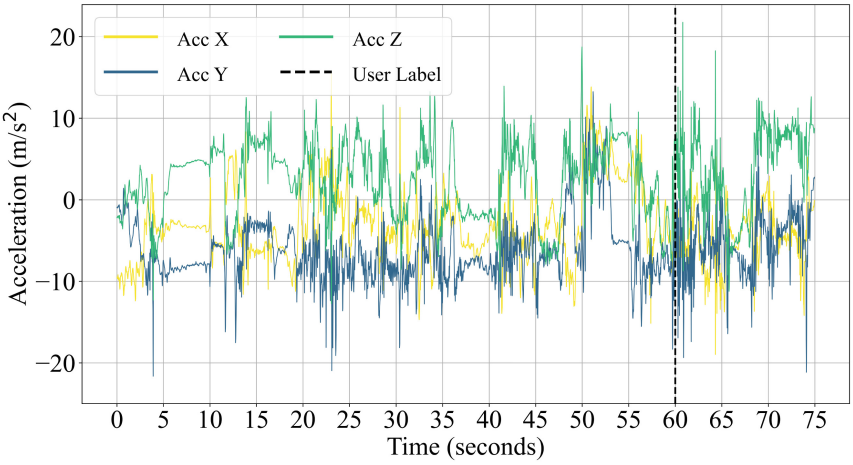
7 Re-labeling Results

To get a first impression of the quality of the manual re-labeling approach, we first create some overall statistics and visualize different aspects of the output.

In Table 1, the number of existing labels for the six different subjects (A-F) as well as the manually set labels (independent from their kind) for the respective annotator are listed. The differences in the number of newly set labels are due to the fact that an annotator could also decide not to set a label at all if he or she believed that there was most likely no activity there. Overall it can be seen that already the amount of user labels differs between the subjects which can be a sign of the severity of the disorder or a general lack of compliance. Subject A clearly shows that a high number of user labels does not mean that handwashing is more routinized and therefore more visually recognizable. Both annotators did not set new labels for almost half of the original user labels. In general, it can be said that the annotators (except for Subject F) were in reasonable agreement as to where handwashing had actually occurred and therefore needed to be labeled.



(a) Example of a handwashing activity where rapid movements are clearly visible. The start of the activity is visually identifiable (around second 8), but the endpoint labeled by the user appears to occur in the midst of a movement.



(b) Example of a handwashing activity where the beginning and end are not easily identifiable due to a less distinctive pattern.

Fig. 1. Visualizations of accelerometer data (in three axes: Acc x, Acc y, and Acc z) depicting two handwashing activities for Subject E. The subject's original labels are represented by dotted vertical lines.

Figure 2 and Fig. 3 give insights into the shares of the different kinds of labels. Figure 2 illustrates the frequency and distribution of label types utilized by the four annotators across all subjects to which they were assigned. This figure provides insights into various annotator behaviors, revealing significant variations in label types despite two annotators consistently relabeling the same subject.

Such discrepancies may stem from a lack of common understanding regarding the identification of handwashing activity patterns or uncertainty regarding which actions constitute handwashing (e.g., drying hands or opening the faucet).

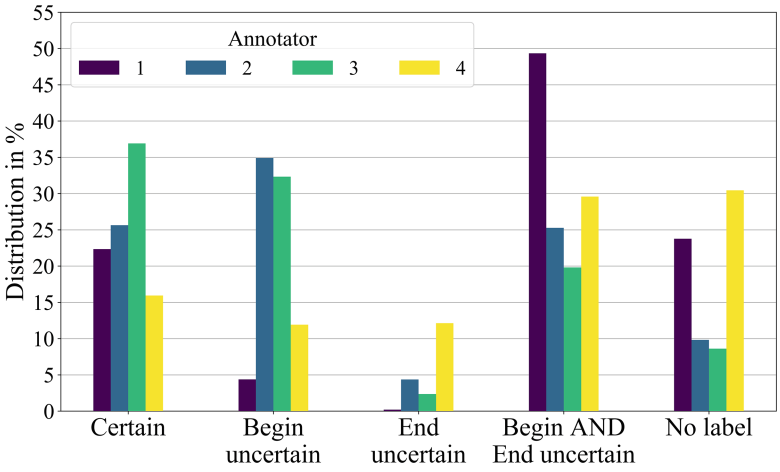


Fig. 2. The figure showcases the spread of different label types assigned by each annotator across all newly set labels.

Figure 3 shows the same label types but in relation to the different participants. This plot aids in identifying subjects where handwashing was visually easier to discern (labeled as *Certain*), as well as instances where the pattern was less clear (labeled as *Begin AND End uncertain*). It also highlights the participants for which incorrect previous user labels may have occurred, potentially indicating no movement and thus no newly set label.

The mean handwashing durations for compulsive as well as routine handwashing in seconds for each subject after re-labeling the data are illustrated in Fig. 4. Therefore, the two annotations for each activity have been combined. It becomes apparent that the duration of handwashing is extremely different not only across participants but also within a participant (recognizable by the high standard deviation). This may be due to natural human behavior, but may also be influenced by the annotators and the ambiguous visual pattern. Since no clear statement can be made about different durations between the two different types of hand washing (compulsive and routine), no conclusions can be drawn here either. Without differentiation between compulsive and routine handwashing, the overall mean duration is 52.30 s with a standard deviation of 39.00 s. With differentiation, the mean durations are 56.09 s for compulsive and 53.84 s for routine behaviors, with standard deviations of 53.89 s and 30.96 s, respectively.

When considering activities where annotators were certain during re-labeling, the frequency of hand washes decreases. Figure 5 visualizes compulsive and rou-

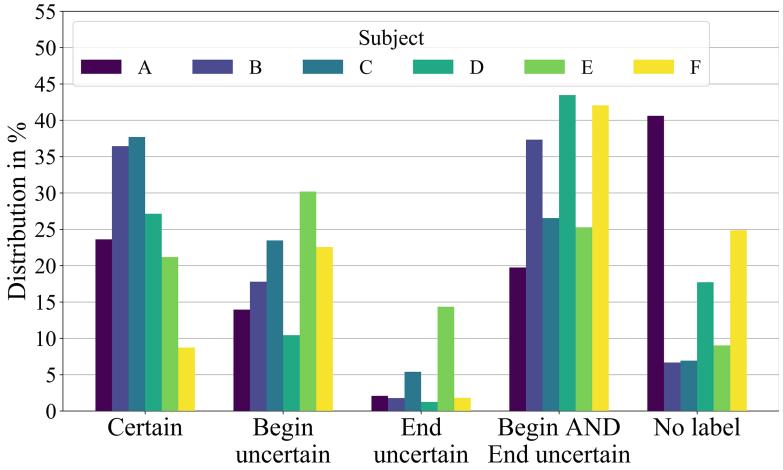


Fig. 3. The visualization depicts the distribution of various label types assigned to different participants.

tine hand washes where both annotators confirmed certainty about the activity pattern. This results in an overall mean duration of 32.02 s with a standard deviation of 13.85 s when not differentiating between handwashing types. When distinguishing between compulsive and routine handwashing, the latter has a mean duration of 36.11 s $\pm$ 17.41 s (with $n = 25$ instances). For $n = 102$ compulsive handwashing activities, the mean duration is 30.48 s $\pm$ 5.92 s. The unexpectedly shorter duration for compulsive handwashing should be interpreted cautiously, as it exhibits a significantly smaller standard deviation compared to routine handwashing activities, despite occurring four times more frequently.

8 IAA Evaluation

As introduced in Sect. 2.2, we used Cohen's Kappa to evaluate the level of agreement between different annotator pairs. The results, visualized in Fig. 6 as a heatmap, reveal considerable variation in agreement levels among the pairs. While annotator pairs $(1, 4)$ and $(3, 4)$ show a higher level of consensus, the other pairs demonstrate lower agreement. It is important to note that these discrepancies are likely to be explained by the different ways of washing hands even within the same participant. Furthermore, it cannot be said with complete certainty that all activities marked by the user were actually handwashing, as even an accidental press of the smartwatch button, for example, would be incorrectly counted as such without the possibility of validating this afterwards.

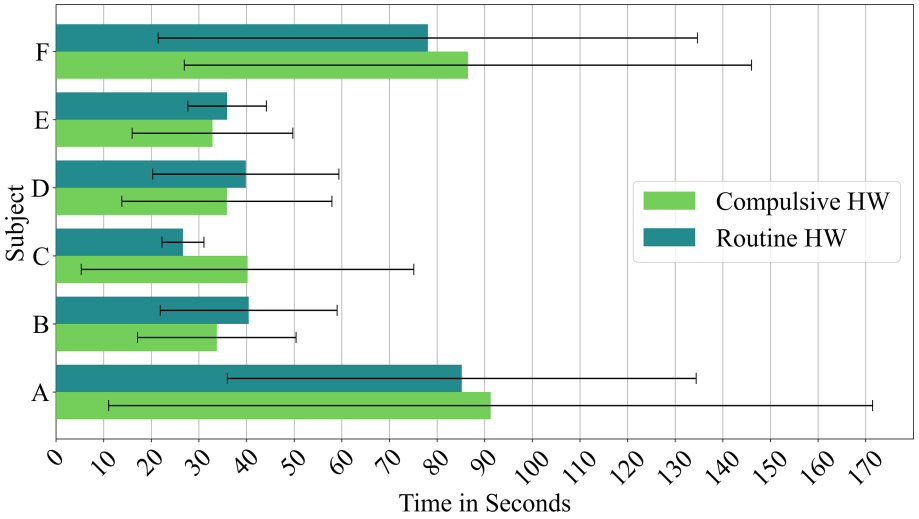


Fig. 4. The horizontal bar chart displays the mean handwashing durations in seconds categorized as compulsive and routine handwashing (abbrev. HW) activities, merged from annotations by each subject. Additionally, each bar also indicates the respective standard deviation.

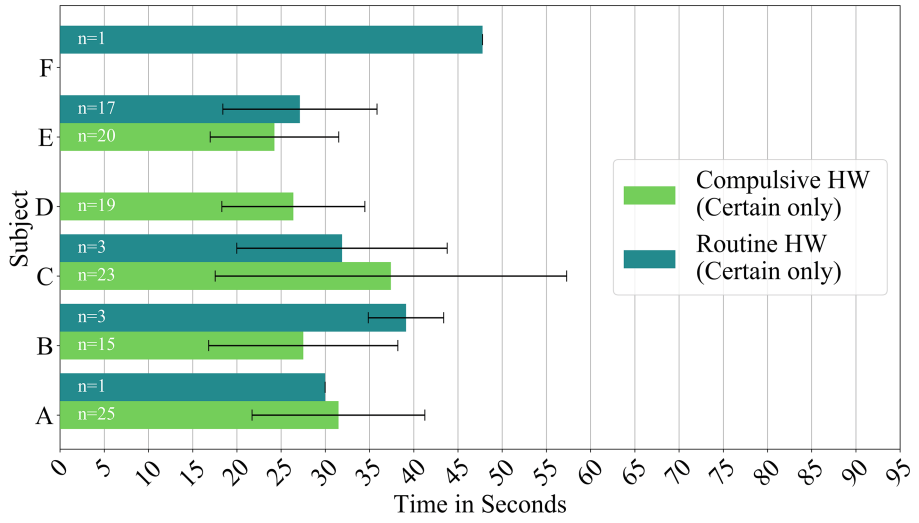


Fig. 5. The horizontal bar chart shows again the mean handwashing durations in seconds categorized as compulsive and routine handwashing activities, merged from annotations by each subject but only when both annotators labeled the activity as being *Certain*. The number of resulting activities is displayed as n if there was at least one.

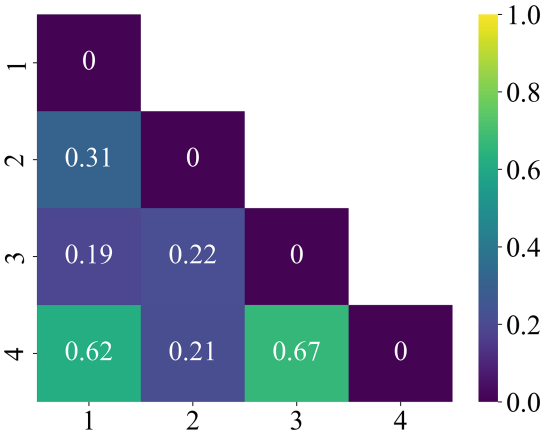


Fig. 6. The heatmap visualizes the IAA by using Cohen's Kappa for each annotator pair.

Cohen's Kappa, commonly used for measuring IAA, has several drawbacks. It is sensitive to an unbalanced distribution of categories, which is particularly relevant in our context. Moreover, it may not accurately reflect agreement when annotators have different tendencies. The simplified treatment of random matches and the sensitivity to annotations near category boundaries further complicate the interpretation. Since in our use case different types of hand washing often follow each other, these factors could lead to lower scores. These limitations underscore the importance of interpreting Kappa scores cautiously.

9 Lessons Learned

As a result of our OCDetect study, we can draw several lessons that are not only valuable for our future work but can also serve the community as a basis for studies in-the-wild.

Plan for Participant Involvement. Following user-centric approach while designing the study is essential. This can involve multiple approaches, ranging from using questionnaires to gather input from the target group about assumptions leading to the study design to incorporating concrete pilot studies as integral parts of the overall research. This might help to understand challenges and increase study results through greater participant commitment and compliance. In our specific case, it would have been helpful to give the participants some (technical) background knowledge on data recording with wearables and maybe even the basic machine learning concept. The lack of understanding of the connection between the time-series data recorded by the smartwatch and the annotation of a handwashing activity by the user led to poorer data quality. Later data exploration showed that some users apparently did not wear the watch, but nevertheless pressed the button to annotate, e.g. after washing their hands. This

behavior is valuable from the user's point of view, as the information for washing hands was provided. From the point of view of automated machine learning, however, this leads to misinformation for the model, as there is no movement data but still a label. In future studies, we will ask users not to set a label if the smartwatch is not being worn. Even if the watch is worn, we will give an approximate time period of 10 min in which an annotation can be made afterwards and, in case of forgetting, simply not to set a label, since a label set much too late can hardly be assigned to the original activity.

Data Recording App Improvements. We have also noticed several times that labels occur in very short succession, where it is unlikely that several activities have taken place. We assume that the user has entered an incorrect label, for example, or that a label has been added several times due to a lack of feedback that a label has already been set. The user interface can be improved technically by introducing the option to take back a label and an overview of annotations that have already been made.

Interdisciplinary Team. In addition, as already demonstrated in our study, it is important to involve not only the user but also experts, such as trained psychologists in our case, in the study process. In this way, trust is created, responsibility is shared and the results can be correctly classified, categorized, and interpreted.

Annotation Guidelines. When it comes to manually annotating the data afterwards, it became apparent that defining annotation guidelines is essential. By doing so, a common understanding of the data and desired outcome is created, the data quality improves considerably, and personal bias is reduced to a minimum.

Multimodal Data Collection. Manual data annotation is time-consuming and requires additional knowledge, such as contextual information. Collecting further data modalities during the study can be beneficial [10]. For activities like handwashing, context, such as location within the room, is crucial. The direct link between specific activities and their spatial allocation (e.g., washing hands at the sink) can help determine the start and end of these activities.

Personalized Pre-model. As previously described in Sect. 5, the smartwatches used for data recording were equipped with a deep learning model pre-trained on lab data. This model aims to automatically recognize as many handwashing activities as possible, requiring confirmation from the user only. However, the study revealed significant variability in individual handwashing techniques, even within the same participant. This variability can lead to numerous incorrect detections, causing label fatigue [12]. To minimize this issue, the pre-trained model can be personalized for each user during the initial days of data recording through a combination of online and active learning with individual data [4]. This personalization reduces false-positives and enhances user compliance.

10 Conclusion

In conclusion, we have presented key findings and challenges in the study design, execution, and data annotation of our in-the-wild study OCDetect. Our findings highlight the importance of a user-centered approach to study design, engaging participants and experts to ensure robust data collection and participant compliance.

Wearable technology proved essential for continuous and unobtrusive data collection in naturalistic settings. However, it also posed a challenge to the accuracy and reliability of the data generated by participants. In our study, both automated and manual relabeling of handwashing activities were performed, showing considerable variability in labeling quality. This highlighted the need for standardized labeling protocols to reduce personal bias and improve data consistency. Although the process of manual relabeling was resource intensive, it provided valuable insights into the reliability of human annotations and IAA. Using Cohen's Kappa metric, we assessed the agreement between annotators.

In a detailed lessons learned section, we highlight the challenges faced during the study and provide potential solutions for future studies of this type, ranging from plans for participant involvement, over additional ideas for data collection and label acquisitions to advanced machine learning approaches.

Acknowledgments. We thank Lorenz Kautzsch and Lea Liekenbrock for their assistance in annotating the data. We would also like to express our gratitude to Silvan Wirth for his technical assistance during data collection and to Alexander Henkel for providing the data recording app.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allahbakhshi, H., Hinrichs, T., Huang, H., Weibel, R.: The key factors in physical activity type detection using real-life data: a systematic review. Front. Physiol. (2019)
2. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in AI safety. arXiv preprint arXiv:1606.06565 (2016)
3. Burchard, R., Scholl, P.M., Lieb, R., Van Laerhoven, K., Wahl, K.: WashSpot: real-time spotting and detection of enacted compulsive hand washing with wearable devices. In: Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing. pp. 483–487. ACM, Cambridge United Kingdom (2022). https://doi.org/10.1145/3544793.3563428
4. Cacciarelli, D., Kulahci, M.: Active learning for data streams: a survey. Mach. Learn. (2024)
5. Caruana, R., Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd International Conference on Machine Learning (2006)

6. Chamberlain, A., Crabtree, A., Rodden, T., Jones, M., Rogers, Y.: Research in the wild: understanding 'in the wild' approaches to design and development. In: Proceedings of the Designing Interactive Systems Conference. DIS 2012, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2317956.2318078
7. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. (1960)
8. Garcia-Ceja, E., Brena, R.F.: Activity recognition using community data to complement small amounts of labeled instances. Sensors (Basel, Switzerland) (2016). https://api.semanticscholar.org/CorpusID:6907811
9. Johnson, R., Rogers, Y., Van Der Linden, J., Bianchi-Berthouze, N.: Being in the thick of in-the-wild studies: the challenges and insights of researcher participation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1135–1144 (2012)
10. Kirsten, K., Arnrich, B.: Elements of a system for automatic monitoring of specific mental health characteristics at home. In: Proceedings of the 25th International Multiconference Information Society 2022, IS 2022, Ljubljana, Slovenia (2022)
11. Komoszyńska, J., Kunc, D., Perz, B., Hebko, A., Kazienko, P., Saganowski, S.: Designing and executing a large-scale real-life affective study. In: 2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), pp. 505–510. IEEE (2024)
12. Kumar, P., Chauhan, S., Awasthi, L.K.: Human activity recognition (HAR) using deep learning: review, methodologies, progress and future research directions. Arch. Comput. Methods Eng. (2024)
13. Larradet, F., Niewiadomski, R., Barresi, G., Caldwell, D.G., Mattos, L.S.: Toward emotion recognition from physiological signals in the wild: approaching the methodological issues in real-life data collection. Front. Psychol. (2020)
14. Mardiko, A.A., von Lengerke, T.: When, how, and how long do adults in germany self-reportedly wash their hands? compliance indices based on handwashing frequency, technique, and duration from a cross-sectional representative survey. Int. J. Hygiene Environ. Health (2020). https://doi.org/10.1016/j.ijheh.2020.113590, https://www.sciencedirect.com/science/article/pii/S1438463920305368
15. Schlögl, S., Buricic, J., Pycha, M.: Wearables in the wild: advocating real-life user studies. In: Proceedings of the 17th International Conference on Human-computer Interaction with Mobile Devices and Services Adjunct (2015)
16. Scholl, P.M., Wahl, K.: Ablutomania-set - a dataset for OCD and everyday handwashing detection from wrist motion (2021). https://earth.informatik.uni-freiburg.de/ablutomania/
17. Tkachenko, M., Malyuk, M., Holmanyuk, A., Liubimov, N.: Label Studio: data labeling software (2020–2022). https://github.com/heartexlabs/label-studio
18. Wahl, K., Scholl, P.M., Miché, M., Wirth, S., Burchard, R., Lieb, R.: Real-time detection of obsessive-compulsive hand washing with wearables: research procedure, usefulness and discriminative performance. J. Obsessive-Compulsive Related Disorders, 100845 (2023). https://doi.org/10.1016/j.jocrd.2023.100845, https://linkinghub.elsevier.com/retrieve/pii/S2211364923000660
19. Wahl, K., et al.: On the automatic detection of enacted compulsive hand washing using commercially available wearable devices. Comput. Biol. Med. 105280 (2022). https://doi.org/10.1016/j.compbiomed.2022.105280, https://linkinghub.elsevier.com/retrieve/pii/S0010482522000725