A Data-Driven Study on the Hawthorne Effect in Sensor-Based Human Activity Recognition

Alexander Hoelzemann*
University of Siegen
Siegen, Germany
alexander.hoelzemann@uni-siegen.de

Marius Bock*
University of Siegen
Siegen, Germany
marius.bock@uni-siegen.de

Ericka Andrea Valladares Bastías
University of Siegen
Siegen, Germany
ericka.vbastias@student.uni-siegen.de

Salma El Ouazzani Touhami
University of Siegen
Siegen, Germany
salma.eotouhami@student.uni-siegen.de

Kenza Nassiri
University of Siegen
Siegen, Germany
kenza.nassiri@student.uni-siegen.de

Kristof Van Laerhoven
University of Siegen
Siegen, Germany
kvl@eti.uni-siegen.de

ABSTRACT

Known as the Hawthorne Effect, studies have shown that participants alter their behavior and execution of activities in response to being observed. With researchers from a multitude of human-centered studies knowing of the existence of the said effect, quantitative studies investigating the neutrality and quality of data gathered in monitored versus unmonitored setups, particularly in the context of Human Activity Recognition (HAR), remain largely under-explored. With the development of tracking devices providing the possibility of carrying out less invasive observation of participants' conduct, this study provides a data-driven approach to measure the effects of observation on participants' execution of five workout-based activities. Using both classical feature analysis and deep learning-based methods we analyze the accelerometer data of 10 participants, showing that a different degree of observation only marginally influences captured patterns and predictive performance of classification algorithms. Although our findings do not dismiss the existence of the Hawthorne Effect, it does challenge the prevailing notion of the applicability of laboratory compared to in-the-wild recorded data. The dataset and code to reproduce our experiments are available via https://github.com/mariusbock/hawthorne_har.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**.

KEYWORDS

human activity recognition, hawthorne effect, data collection

*Authors contributed equally to this research.

UbiComp/ISWC '23 Adjunct, October 08–12, 2023, Cancun, Quintana Roo, Mexico

ACM ISBN 979-8-4007-0200-6/23/10…$15.00
https://doi.org/10.1145/3594739.3610743

ACM Reference Format:
Alexander Hoelzemann, Marius Bock, Ericka Andrea Valladares Bastías, Salma El Ouazzani Touhami, Kenza Nassiri, and Kristof Van Laerhoven. 2023. A Data-Driven Study on the Hawthorne Effect in Sensor-Based Human Activity Recognition. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing (UbiComp/ISWC '23 Adjunct), October 08–12, 2023, Cancun, Quintana Roo, Mexico.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3594739.3610743

1 INTRODUCTION

Body-worn sensor systems bare great potential in analyzing our daily activities with minimal intrusion, yielding various applications from the provision of medical support to supporting complex work processes [3]. With (deep) neural networks representing the state-of-the-art technology for the automatic analysis of such wearable data, a bottleneck becomes the correct annotation of data for the underlying training. Due to the fact that inertial data is difficult to interpret in hindsight without any additional context, most publicly available datasets remain captured in controlled, video-recorded environments with researchers being in close proximity to study participants.

The Hawthorne Effect, originally discovered in 1958 [9], describes the phenomenon that humans alter their behavior and execution of activities in response to being observed. The phenomenon's discovery has led to numerous studies trying to measure the said effect in clinical trials [1, 10, 12, 15, 16, 19], and, more targeted toward physical activities, showed that the effects can be quantified, for instance with gait parameters like step length and cadence of gaits [19]. With researchers from a multitude of human-centered studies being aware of the existence of such an effect, a data-driven study of the phenomenon and its potential effects remain largely under-explored in the community of Human Activity Recognition (HAR). Given that the performance and applicability of learning algorithms, such as neural networks, in real-world scenarios heavily depend on the representativeness of the training data, our study aims to investigate whether the prominent observation of participants during data collection introduces biases and results in measurable and altered executions of activities which, in turn,

may potentially lead to less effective and less generalized networks. Inspired by the works of Vickers *et al.* [19], our paper provides a data-centric analysis of measuring a possible Hawthorne Effect on a variety of fitness activities through the modality of wrist-worn inertial sensor data. This is done by explicitly letting the participants be observed through cameras and/or the researchers. Contributions of our paper are three-fold:

(1) We designed a HAR experiment where volunteers perform a set of activities under three observation settings: 1) fully-observed (video-recorded + monitoring by researchers), 2) semi-observed (video-recorded + no monitoring), and 3) non-observed (no video recording + no monitoring).
(2) We collected data from 10 participants performing 5 different activities, *jumping, walking, jogging in place, sit-ups, and jumping jacks* over several days.
(3) We perform both feature analysis and investigation of changes in the predictive performance of a deep learning classifier [2] based on the type of observation applied during validation as well as its capabilities to distinguish between each participant's session.

2 RELATED WORK

Human Activity Recognition typically relies on participants being monitored via wearable sensors, making them consistently aware of being observed. However, these circumstances may have introduced a behavior bias [20] into publicly available datasets. This bias manifests as changes in behavior when study participants are aware of being monitored by another person or a video recording system [7] and is also known as the **Hawthorne Effect**. The fundamental research of which the Hawthorne effect originated, was conducted between 1924 and 1927 as part of an investigation of whether the productivity of workers of the Hawthorne Western Electric plant could be increased by a change in lighting conditions [11]. With later studies criticizing the research methodology [4], Landsberger concluded in 1958 [9] that the increase in productivity was to be attributed to the workers being aware that they were monitored and not the change in working conditions. The observed phenomenon, i.e. the alteration of behavior whenever participants are aware that they are being monitored, was later then primed as the Hawthorne Effect. In the context of HAR experiments, the Hawthorne Effect suggests that participants' awareness of being monitored can potentially affect the applicability and generalization of trained activity recognition systems. Knowing they are being observed and activities are being recorded, participants might result in modifying their movements, behaviors, and/ or daily routines, leading to a deviation from a natural execution of activities. This alteration can introduce biases and inaccuracies in the data collected for HAR experiments, making it challenging to develop reliable and scalable activity recognition systems [4]. To mitigate the Hawthorne Effect in HAR experiments, researchers often opt for minimizing participants' awareness of being monitored. By employing discrete sensing techniques, such as using a minimalistic setup of wearable devices [17] or ambient sensors [13], they can collect activity data without participants constantly focusing on the monitoring process. By reducing the conscious attention given to monitoring, researchers aim to capture more natural and representative data that can improve the accuracy and reliability of HAR systems [8].

In 2014, Bulling *et al.* [3] described several research challenges in creating datasets for human activity recognition that avoid bias. These challenges, being still relevant to this date, include *intra-class* variability, *inter-class* similarity, and the *NULL-class* problem. Important in the context of the Hawthorne Effect is the *intra-class* variability, which describes how data from the same class differ between participants or sometimes even instances of one activity from the same individual due to stress, fatigue, or an emotional or environmental state in which the activity is performed. As such, the Hawthorne Effect can be categorized as an *intra-class* variability problem - which can have a direct effect on classifier capabilities and performance.

3 METHODOLOGY

Study Protocol: To investigate any potential effects observation of participants can have on collected inertial data, we asked 10 participants (4 females, 6 males) to perform a short workout across multiple days, employing different types of observation (see Figure 1). The study was approved by our university's ethics council. Study participation was voluntary, and informed consent was obtained from all participants before the study. The workout plan consisted of a fixed order of 5 different activities, i.e. *jumping*, *walking*, *jogging_in_place*, *sit_ups*, and *jumping_jacks*, each performed for 120 seconds with breaks in-between the activities.

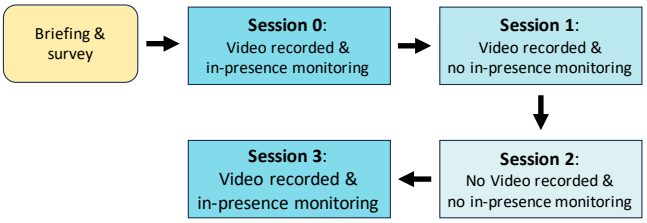

Figure 1: Applied study protocol. After having a briefing and filling out a pre-study survey, each participant performed the workout 4 times across 4 different days. The first and last workouts (sessions 0 and 3) were video-recorded and performed under the observation of at least one researcher. The second and third workouts (sessions 1 and 2) were performed without the observation of any researcher at a location chosen by the participant (e.g. at home). The second workout (session 1) was additionally video-recorded.

Before their first workout session, each participant was briefed about the study protocol in a structured manner and shown sample data collected by the tracking device. Participants were further asked to answer a short survey asking for gender and age group as well as whether they perform regular private workouts and a fitness tracker in their daily lives (see Table 1). To avoid any unwanted biases, participants remained unaware throughout the study that the data would be analyzed to assess differences between supervised and unsupervised recording setups. In total, each participant was tasked to perform the workout 4 times using 3 different types of observation. After having been briefed, each participant was equipped with a smartwatch on their wrist of choice and given a demonstration by one of the researchers of the correct execution of each exercise. The smartwatch, a Bangle.js Version 1, was set

Table 1: Pre-study survey answers provided by each participant. The survey asked participants to provide age, gender, and whether they perform regular private workouts and wear a wearable device (e.g. fitness smartwatch) in their daily lives.

ID	Age	Gender	Pvt. Workouts	Pvt. Wearable
sbj_0	18-25	F	✗	✗
sbj_1	26-35	F	✓	✗
sbj_2	26-35	M	✓	✓
sbj_3	18-25	M	✓	✓
sbj_4	18-25	F	✗	✓
sbj_5	26-35	M	✓	✓
sbj_6	18-25	F	✗	✓
sbj_7	18-25	M	✗	✗
sbj_8	26-35	M	✗	✗
sbj_9	26-35	M	✗	✗

to record 3D accelerometer data at a constant sampling rate of 12.5 Hz with a sensitivity of ±8g using a custom, open-source firmware[18]. The first workout session (*session 0*) was performed under the observation of one researcher in a location decided by the participant and researchers with a video-recording device taping the execution of the routine. After the completion of *session 0*, participants were walked through the control of the smartwatch and tasked to perform the workout plan within the next days twice at a location of their choice (e.g. their home) – one-time video-recording (*session 1*) and another time without video-recording their workout (*session 2*). Lastly, participants were invited back to where they originally performed the first session and asked to perform the workout a second time under the observation of a researcher with a video recording in place (*session 3*). In between the 4 sessions participants were asked to wear the smartwatch as much as possible throughout their daily life, keep a brief recount of their daily activities and note down the start and end times of each of them. To further ensure the workout of interest can properly be identified in the activity streams, each session started and ended with the activity *jumping*. Having identified the workouts in the inertial data recordings, the 3D-acceleration data streams were cropped to only include the workout activities, labeled accordingly, and saved session-wise for each participant into separate files.

Feature Analysis. The feature analysis incorporates a Fast Fourier Transform (FFT), as depicted in Figure 2 and a comparison of the total number of repetitions, indicated by the $\sum$-sign, and repetitions per second for a specific activity, indicated by the $\varnothing$-sign, taking into account the subject and session in which the activity was performed. The results are presented in Table 2. Both, the FFT [6] and the repetitions per second, calculated with a peak detection algorithm [5], are calculated utilizing functions provided by the SciPy community. The peak detection algorithm specifically processes 1-dimensional time-series data. For our analysis, we computed the magnitude of the accelerometer signal and employed it for the algorithm. Given the periodic nature of the activities under study, each positive peak observed in the signal can be interpreted as indicating one repetition. To validate the accuracy of the peak detection, a visual confirmation was conducted.

Deep Learning Analysis. As proposed by Ordoñez and Roggen in [14], a popular methodology in human activity recognition remains the usage of convolutional and recurrent layers. The former is used to automatically extract discriminative features. Having shown quantitative differences in the feature analysis, the following will investigate the effect said (potential) differences have on the performance and applicability of neural networks. All experiments were conducted using a shallow DeepConvLSTM [2] employing a kernel size of 3, 1024 hidden LSTM units, and inertial data which was split into sliding windows of 1 second with a 50% overlap. We reuse hyperparameters reported in [2], proven to work on a multitude of activity recognition datasets, and only increase the number of epochs (300) while employing a step-wise learning rate schedule, decreasing the learning rate by a factor of 0.9 every 30 epochs. To minimize the risk of performance differences between experiments being the result of statistical variance, reported metrics are averaged across three runs using three different random seeds. Our three types of experiments aim at answering three types of questions: **(1) Cross-session generalization:** How well does a network, trained using fully-observed data, predict data recorded employing different degrees of observation? That is, for each subject, predict each session's activities individually having trained on all other subjects' session 0 data. **(2) Session differentiation:** Can a network be trained to classify data records into the respective session type they originate from? That is, for each subject, predict each data records session type having trained on all other subjects' data. **(3) Fully-observed overfitting:** Are patterns learned by a network overfitted on a subject's fully-observed data transferable to data employing different degrees of observation? That is, for each subject, predict activities recorded during sessions 1, 2, and 3 using a network overfitted, i.e. reaching close-to-perfect classification scores, on session 0 data. In order to achieve the network overfitting, these experiments involve increasing the number of epochs (1000), learning rate (0.2) and applied learning rate scheduler step size (250).

4 RESULTS

Overall, we were not able to prove that the Hawthorn Effect is directly verifiable by any of the aforementioned analyzing methods or that data recorded in various recording environments (controlled or semi-controlled) differ significantly.

Feature Analysis. The analysis of the Fast Fourier Transform, Figure 2, reveals that fully monitored sessions 0 and 3 generally do not exhibit similar dominant frequencies, which is also the case for semi-monitored and unmonitored sessions 1 and 2.

In particular, several activities and subjects align with our previously established hypothesis that the signal from session 3 should converge back to that of session 0. Such behavior is evident for *sbj_6*, *sbj_7* and *sbj_8* during the *jumping_jacks* activity, and for *sbj_2* during the *sit_ups* activity. It is important to acknowledge that this bias may have arisen both due to the researcher's observations and the spatial variations in the workout environment. The fact that this behavior is more evident while executing *jumping_jacks*, might indicate that a Hawthorne Effect is limited to a specific kind of activity. Further noteworthy differences are visible in the signals of the activities *jumping*, *jogging_in_place* and *sit_ups*. Here, the results of the FFT suggest that *sbj_1*, *sbj_4*, and *sbj_9* altered their activity execution behavior depending on the experience gained during the study. The light-green (session 0) and blue (session 1) most dominant frequencies are more similar to each other than red

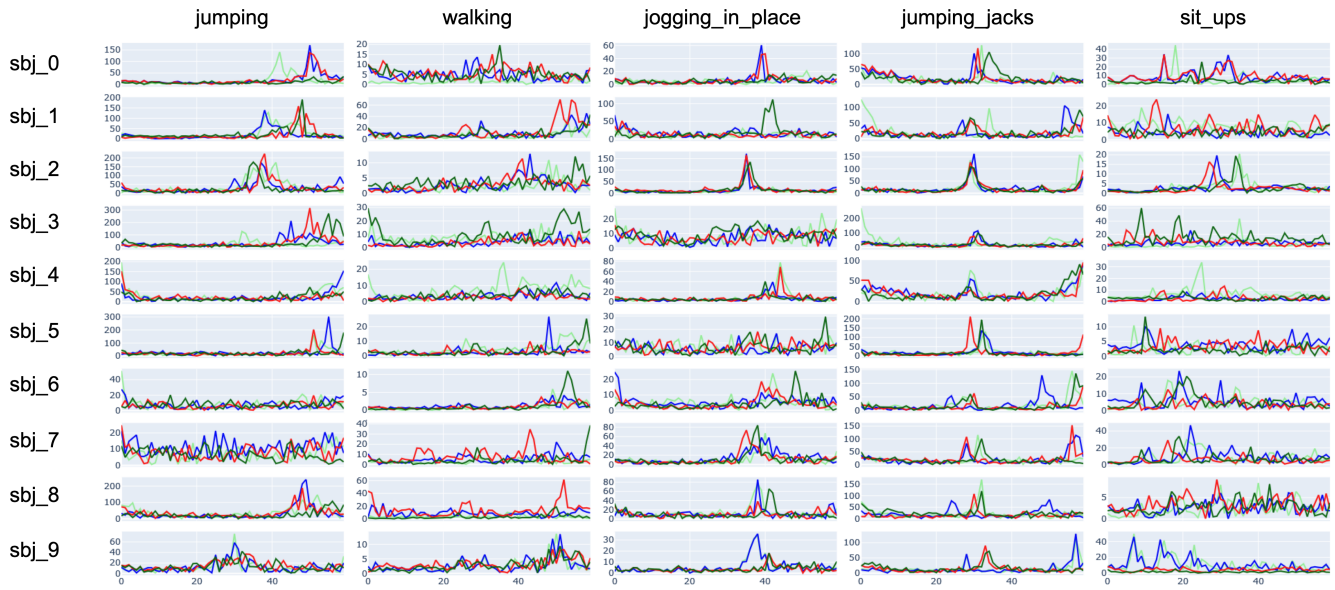

Figure 2: Fast Fourier Transform (FFT) calculated on every participant and every activity included in our study. Light-green represents session 0 (monitored and video recorded), blue session 1 (non-monitored, video recorded at home), red session 2 (non-monitored, non-video-recorded), dark-green session 3 (monitored and video recorded)

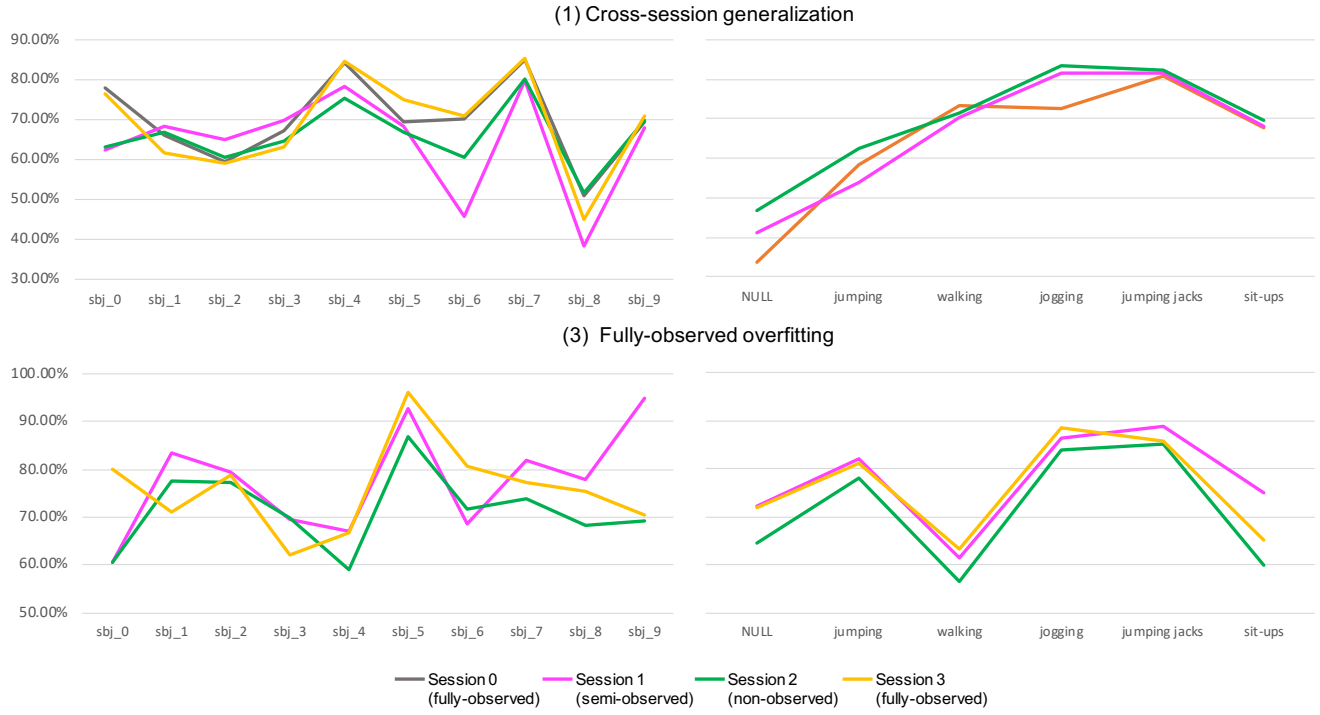

Figure 3: Per-subject and per-activity accuracy results of the (1) cross-session generalization and (3) fully-observed overfitting experiments. Results are averaged across three runs using three different seeds. With the exception of *sbj_6* differences amongst sessions remain marginal. Though producing the on-average lowest results, data recorded in semi- and non-observed environments shows to be similarly applicable in terms of predictive performance than compared to fully-observed data, and, in the case of the semi-observed data, is even more reliably predicted by a network overfitted on fully-observed data.

Table 2: This table shows the total number of repetitions (∑) encountered in the activities' signal and the number of repetitions per second (∅). Every subject has 4 rows that represent the session. Activities that are marked as ▭ represent cases where the number of repetitions per second was higher during the monitored sessions than during the unmonitored ones, activities colored in ▭ are activities where the number of repetitions per seconds was higher during the unmonitored sessions than during the monitored sessions.

	jumping ∑ , ∅	walking ∑, ∅	jogging_in_place ∑, ∅	jumping_jacks ∑, ∅	sit_ups ∑, ∅
sbj_0	41, 1.68	39, 1.77	79, 1.92	32, 1.01	20, 0.61
	51, 1.66	41, 1.28	43, 1.44	31, 1.03	16, 0,53
	51, 1.64	40, 1.11	43, 1.41	32, 1.03	16, 0.52
	61, 1.85	54, 1.34	75, 2.02	35, 1.08	27, 0.75
sbj_1	42, 1.33	45, 1.48	68, 2.17	65, 1.75	9, 0.23
	40, 1.32	58, 1.70	60, 2.04	49, 1.66	10, 0.30
	48, 1.48	51, 1.73	60, 2.08	47, 1.50	14, 0.27
	47, 1.55	56, 1.55	39, 2.08	54, 1.73	11, 0.35
sbj_2	19, 0.70	48, 1.28	36, 1.15	30, 0.97	16, 0.37
	19, 0.62	41, 1.34	37, 1.19	31, 1.00	13, 0.40
	22, 0.69	39, 1.42	36, 1.20	31, 1.03	10, 0.32
	18, 0.60	53, 1.33	36, 1.14	31, 0.99	16, 0.44
sbj_3	35, 1.08	67, 1.28	80, 2.62	40, 1.14	14, 0.43
	32, 1.09	50, 1.35	78, 2.68	41, 1.36	13, 0.41
	27, 0.96	37, 1.27	71, 2.61	42, 1.46	6, 0.364
	30, 1.03	48, 1.44	72, 2.37	37, 1.20	10, 0.35
sbj_4	31, 1.07	39, 1.33	45, 1.45	58, 2.01	13, 0.45
	30, 1.07	40, 1.14	42, 1.40	60, 1.93	10, 0.30
	35, 1.16	37, 1.30	44, 1.45	59, 1.91	13, 0.38
	29, 1.10	40, 1.18	47, 1.52	55, 1.90	10, 0.29
sbj_5	46, 1.59	46, 1.37	31, 1.03	33, 1.02	8, 0.32
	47, 1.50	44, 1.42	32, 0.95	33, 1.05	10, 0.31
	50, 1.68	35, 1.16	30, 0.98	31, 1.09	13, 0.26
	46, 1.35	48, 1.27	28, 0.99	33, 1.05	11, 0.36
sbj_6	40, 1.25	52, 1.42	83, 2.58	55, 1.63	11, 0.34
	42, 1.30	48, 1.35	78, 2.46	49, 1.63	11, 0.31
	46, 1.44	42, 1.35	81, 2.58	48, 1.55	8, 0.244
	52, 1.52	52, 1.56	87, 2.52	57, 1.75	11, 0.32
sbj_7	50, 1.58	42, 1.14	55, 1.64	31, 0.93	12, 0.35
	56, 1.76	36, 1.16	40, 1.34	27, 0.89	12, 0.31
	54, 1.79	27, 1.04	36, 1.18	28, 0.95	9, 0.30
	56, 1.87	40, 1.09	39, 1.22	31, 0.94	11, 0.34
sbj_8	23, 0.76	36, 1.34	36, 1.21	31, 0.95	11, 0.34
	29, 0.96	40, 1.32	33, 1.10	21, 0.87	10, 0.33
	25, 0.90	40, 1.28	33, 1.07	24, 0.81	10, 0.32
	53, 1.85	45, 1.46	33, 1.06	23, 0.74	9, 0.29
sbj_9	27, 1.06	48, 1.45	66, 2.34	37, 1.41	10, 0.28
	27, 1.05	48, 1.44	67, 2.37	37, 1.42	8, 0.23
	25, 0.91	49, 1.37	65, 2.59	31, 1.08	10, 0.34
	24, 0.92	48, 1.38	65, 2.60	32, 1.08	9, 0.33

(session 3) and dark-green (session 4), which in turn show greater similarities to each other than compared to the first two sessions 0 and 1.

Table 2 provides a color-coded depiction of repetition patterns across each individual recording session. One can see that the table does not reveal any universally applicable patterns that confirm the Hawthorne Effect across all scenarios. Only two subjects, *sbj_0* and *sbj_2*, demonstrate a difference in the number of repetitions per second between monitored and unmonitored sessions. More specifically, *sbj_0* shows an increase in repetitions for 4 out of 5 activities when observed by a researcher, with only the activity *jumping_jacks* not showing such a trend. However, this activity shows an equal number of repetitions per second for both unsupervised sessions. Similarly, *sbj_2* demonstrates a higher frequency of repetitions for the activities *walking*, *jogging_in_place*, and *jumping_jacks*; yet, this behavior is even less commonly observed compared to the first scenario.

Deep Learning Analysis. Table 3 summarizes the average accuracy and macro F1-scores obtained during each of the three deep-learning-based experiments. Using data recorded during session 0, i.e. data originating from the same session as data used for training the network, resulted, as expected, in the highest validation metrics (70% accuracy and 64% macro F1-score). Further, being recorded under the same conditions, validating using data recorded during session 3 resulted in the second-to-best results, being on average only around a percentage point worse than the validation using session 0 data. Surprisingly, using the self-recorded participant data (sessions 1 and 2) for validation did not result in a significant drop in performance. Even though participants recorded themselves in a completely unmonitored recording setup (session 2) performance drops were only around 4% compared to using fully-observed data. With accuracy scores being close to random guessing, Table 3 further shows that the shallow DeepConvLSTM [2] was incapable of being trained to differentiate data records based on the session which they originate from. Lastly, inference of networks overfitted on session 0 data showed to produce similar results across all sessions, with, though applying a different observation scenario, session 1 (semi-observed) producing the highest classification results. Overall, the results of all three experiments suggest that the predictive performance of the network of choice only marginally suffers when being used for inference on data recorded by applying a different degree of observation.

Table 3: Average accuracy and F1-scores of the three types of performed experiments ((1), (2) and (3)). Experiments are divided by the type of session data used during validation. Results are the averages and standard deviation across subjects across three runs using three different seeds. Note that given the altered prediction scenario (session instead of activity type) experiment (2) does not involve splitting each subject's data into different session types.

Exp	Val. Set	Accuracy	F1-score
(1)	Session 0	70.11 ± 10.55	64.45 ± 12.07
	Session 1	64.46 ± 13.11	60.18 ± 13.48
	Session 2	66.02 ± 8.10	62.35 ± 8.93
	Session 3	69.33 ± 12.34	65.81 ± 12.45
(2)	All	25.21 ± 3.32	21.90 ± 3.05
(3)	Session 1	77.58 ± 11.18	76.74 ± 11.37
	Session 2	71.42 ± 8.21	69.55 ± 9.89
	Session 3	75.87 ± 9.31	74.27 ± 10.32

Especially visualization of the per-class and per-participant results of the *(1) cross-session generalization* and *(3) fully-observed overfitting* experiments (see Figure 3) shows that, besides *sbj_6*, results remained stable across all sessions. Even though the non-observed setup (session 3) remained on average the least performant session, it nevertheless shows the lowest standard deviation across subjects.

5 CONCLUSIONS AND DISCUSSION

This paper presented a data-driven investigation aiming at measuring the effects of the Hawthorne Effect in the context of Human Activity Recognition. The study involved the recording of 10 participants performing 5 distinct activities on 4 different days. With the first and last day being supervised and video-recorded by researchers, the remaining two days had participants self-record themselves at a location of their choice with and without video recording in place. To avoid potential biases, participants remained unaware throughout the study that the data would be analyzed to assess differences between supervised and unsupervised sessions. As part of analyzing the captured data, we employed a feature and deep learning analysis, ultimately concluding that the recorded data does not exhibit a measurable Hawthorne Effect. While the feature analysis did not reveal any generalizable patterns, the deep learning analysis showed that data originating from the unmonitored sessions produced similar classification results and even outperformed the fully observed in some cases. Although our findings do not dismiss the existence of the Hawthorne Effect, especially given the numerous clinical trials proving said effect, (see Section 1), it does challenge the prevailing notion of the applicability of laboratory compared to in-the-wild recorded data. Results of our study show that though an altered behavior of participants might be present, classification algorithms seem to learn discriminative features of similar applicability regardless.

At this point, it is important to acknowledge the limitations of this study and discuss possible reasons why the effects between the different observation scenarios were not as pronounced as we hypothesized when designing the study. Generally, the recorded dataset may not have the necessary size to draw generalizable conclusions and can only indicate a trend. Furthermore, the recorded activities solely represent a subset of periodic nature within the broader context of activity recognition. Several reasons for the lack of significant differences could be: **(1)** The participants were for all observation settings made aware that their inertial data was recorded (as this is required by the ethics council). This might mean that a possible Hawthorne Effect could have been present under all measured conditions and that this was not more pronounced when observed by additional cameras and the researchers being present. **(2)** The choice of activities could have resulted in overly simplistic movement classes that make it hard to find stark differences between the different observation sessions through our analysis methods. **(3)** It is also possible that the Hawthorne Effect in general is relatively small for our five-activity-class scenario when compared to more behavior-oriented activities (such as "brushing teeth") or fine-grained characterizations (for instance for gait analysis).

Due to the inherently limited interpretability of neural networks and the opaqueness of their decision-making processes, it is uncertain whether the observed disparities in prediction performance can be attributed solely to varying learned feature representations resulting from different levels of observation. Therefore, further investigation is warranted to explore the presence and potential impact of these influences. We believe that the results of this paper's study are nevertheless worthy of more discussion and we encourage others to perform further, more extensive research on this topic.

ACKNOWLEDGMENTS

Marius Bock is funded by the DFG Project WASEDO (DFG LA 275811-1).

REFERENCES

[1] Fabrizio Benedetti, Elisa Carlino, and Alessandro Piedimonte. 2016. Increasing uncertainty in CNS clinical trials: the role of placebo, nocebo, and Hawthorne effects. *The Lancet Neurology* 15, 7 (2016). https://doi.org/10.1016/S1474-4422(16)00066-1

[2] Marius Bock, Alexander Hoelzemann, Michael Moeller, and Kristof Van Laerhoven. 2021. Improving Deep Learning for HAR with Shallow LSTMs. In *ACM International Symposium on Wearable Computers*. https://doi.org/10.1145/3460421.3480419

[3] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *Comput. Surveys* 46, 3 (2014). https://doi.org/10.1145/2499621

[4] Hao Chen, Seung H. Cha, and Tae W. Kim. 2019. A framework for group activity detection and recognition using smartphone sensors and beacons. *Building and Environment* 158 (2019). https://doi.org/10.1016/j.buildenv.2019.05.016

[5] The SciPy community. 2023. Calculate the relative extrema of data. https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.argrelextrema.html, Last accessed on 2023-06-29.

[6] The SciPy community. 2023. Fast Four Transformation. https://docs.scipy.org/doc/scipy/tutorial/fft.html, Last accessed on 2023-06-29.

[7] Kenzie B Friesen, Zhaotong Zhang, Patrick G Monaghan, Gretchen D Oliver, and Jaimie A Roper. 2020. All eyes on you: how researcher presence changes the way you walk. *Scientific Reports* 10, 1 (2020), 1–8.

[8] Alexander Hoelzemann and Kristof Van Laerhoven. 2023. A matter of annotation: An empirical study on in situ and self-recall activity annotations from wearable sensors. *CoRR* abs/2305.08752 (2023). https://arxiv.org/abs/2305.08752

[9] Henry A. Landsberger. 1958. Hawthorne revisited: Management and the worker, its critics, and developments in human relations in industry. *Cornell Studies in Industrial and Labor Relations* 9 (1958). https://doi.org/10.1086/222616

[10] Connor Malchow and Goeran Fiedler. 2016. Effect of Observation on Lower Limb Prosthesis Gait Biomechanics: Preliminary Results. *Prosthetics and Orthotics International* 40, 6 (2016). https://doi.org/10.1177/0309364615605374

[11] Elton Mayo. 1930. The human effect of mechanization. *The American Economic Review* 20, 1 (1930). https://www.jstor.org/stable/1805670

[12] Jim McCambridge, John Witton, and Diana R. Elbourne. 2014. Systematic Review of the Hawthorne Effect: New Concepts Are Needed to Study Research Participation Effects. *Journal of Clinical Epidemiology* 67, 3 (2014). https://doi.org/10.1016/j.jclinepi.2013.08.015

[13] Ghassem Mokhtari, Qing Zhang, Ghavameddin Nourbakhsh, Stephen Ball, and Mohanraj Karunanithi. 2017. BLUESOUND: A new resident identification Sensor - Using ultrasound array and BLE technology for smart home platform. *IEEE Sensors Journal* 17, 5 (2017). https://doi.org/10.1109/JSEN.2017.2647960

[14] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* 16, 1 (2016). https://doi.org/10.3390/s16010115

[15] Edward Purssell, Nicholas Drey, Jane Chudleigh, Sile Creedon, and Dinah J. Gould. 2020. The Hawthorne effect on adherence to hand hygiene in patient care. *Journal of Hospital Infection* 106, 2 (2020). https://doi.org/10.1016/j.jhin.2020.07.028

[16] Verónica Robles-García, Yoanna Corral-Bergantiños, Nelson Espinosa, María Amalia Jácome, Carlos García-Sancho, Javier Cudeiro, and Pablo Arias. 2015. Spatiotemporal Gait Patterns During Overt and Covert Evaluation in Patients With Parkinson's Disease and Healthy Subjects: Is There a Hawthorne Effect? *Journal of Applied Biomechanics* 31, 3 (2015). https://doi.org/10.1123/jab.2013-0319

[17] Yonatan Vaizman, Nadir Weibel, and Gert Lanckriet. 2018. Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. *ACM Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018). https://doi.org/10.1145/3161192

[18] Kristof Van Laerhoven, Alexander Hoelzemann, Iris Pahmeier, Andrea Teti, and Lars Gabrys. 2022. Validation of an open-source ambulatory assessment system in support of replicable activity studies. *German Journal of Exercise and Sport Research* 52, 2 (2022). https://doi.org/10.1007/s12662-022-00813-2

[19] Joshua Vickers, Austin Reed, Robert Decker, Bryan P. Conrad, Marissa Olegario-Nebel, and Heather K. Vincent. 2017. Effect of Investigator Observation on Gait Parameters in Individuals With and Without Chronic Low Back Pain. *Gait & Posture* 53 (2017). https://doi.org/10.1016/j.gaitpost.2017.01.002

[20] Kristina Y. Yordanova, Adeline Paiement, Max Schröder, Emma Tonkin, Przemyslaw Woznowski, Carl Magnus Olsson, Joseph Rafferty, and Timo Sztyler. 2018. Challenges in Annotation of useR Data for UbiquitOUs Systems: Results from the 1st ARDUOUS Workshop. *CoRR* abs/1803.05843 (2018). http://arxiv.org/abs/1803.05843