Towards a Methodology for Acceptance Testing and Validation of Monitoring Bodyworn Devices

Emma L. Tonkin, Miquel Perello Nieto, Haixia Bi, Antonis Vafeas
Digital Health
School of Electrical and Computer Engineering
The University of Bristol, Bristol, UK
Email: {e.l.tonkin,miquel.perellonieto,haixia.bi,antonis.vafeas}@bristol.ac.uk

***Abstract*—In hardware deployments, it is often necessary to test platforms for suitability for particular purposes. As lengthy data collection processes often outlive specific iterations of hardware and firmware, it is likely that migration between platforms may become necessary. In this short paper we describe a practical approach employed for acceptance testing, comparison and validation of two iterations of a wearable accelerometer and localisation platform, based on an annotated 15-task activity script. We present an analysis on the data generated for the different activities, and compare device performance using common machine learning algorithms for activity recognition.**

I. Introduction

Wrist-worn wearable sensor devices are increasingly used for the purpose of monitoring health-related metrics. A variety of sensors may be used for these purposes. For example, accelerometers may be used to evaluate motion or characterise activity [1], whilst the strength of radio-frequency signals sent by a wearable device may be used to triangulate the wearer within the home, thus providing localisation information [2]. Biosensors may be used to monitor metrics such as pulse rate, blood pressure, etc. [3].

For sensors of this nature to be used as an input to clinical practice, a process of sensor validation is recommended. On the most basic level this means that the characteristics of the sensor output must be established and characterised [4]. Within an application – for example, monitoring a patient for a given clinical purpose – validation may further imply establishing the extent of any equivalence between the measures given by the sensor alongside any further analysis that takes place on the data and the clinical metrics of interest to the clinician.

This validation process can be complex and lengthy, especially in applications requiring active engagements with patients. Once completed, the cost of this process is potentially a limiting factor in the decision to modify devices (for example, alter the sensors used in wearable devices or add features to firmware). Yet there are many good reasons for hardware, firmware and software within wearable devices, such as the need to add features, to alter components as old ones become available or better alternatives come onto the market, to reduce power use and lengthen battery life, or to resolve bugs identified during deployment. Hence, it is useful to explore means by which modified versions of devices may be tested for functional equivalence within a given frame of reference (i.e. use case).

Note that this testing approach is complemented by a 'burn-in' test completed prior to the deployment of each individual set of devices. Individual devices may fail early in their lifespan – it is often observed in quality analysis that device failures typically occur either early in their useful life (for example, due to manufacturing defects), or as a result of age-related wear, although the precise distribution may vary [5]. To capture this type of problem, individual devices are tested for approximately seven days prior to deployment. The process described here, however, relates to acceptance testing of a class of devices rather than of individual devices within that group.

A. Evaluating data quality by comparing sensors

In this paper we compare two wrist worn wearable sensors. In particular, the wearable sensors are two successive versions of the SPHERE wearable, which includes a three-axis accelerometer and a bluetooth low-energy (BLE) based localisation mechanism. We compare the quality of data generated from two versions of the sensors. The data is associated with wrist movement via a 3-axis accelerometer and room level location information extracted from receivers picking-up packets from SPHERE wearable devices via BLE.

These devices are used in long-term free-living observation studies, such as the SPHERE 100-Homes study, and in studies monitoring patient recovery following hip and knee replacement surgery. They are also in use with participants with Alzheimers and Parkinsons, and their use for sleep monitoring alongside EEG monitoring has been explored in a recent study. Our primary interest in these contexts is in participants' ways of living 'in the wild'. Much of the work done focuses on the use of data as a comparable proxy to clinical datasets. For example, for the purpose of establishing the extent to which a patient participates in everyday activities, we may focus on quality of motion, ease of completion of tasks such as sit-to-stand or stairclimbing, and detection of everyday activities such as making a cup of tea.

For the purpose of acceptance testing of a new version, our focus is not on absolute quality of data. That is established in the validation processes themselves. Instead, we are interested in similarity of or improvement in of the new platform's performance, compared with the version that was previously in use and under evaluation. In the context of healthcare, we require that a new version is able to capture enough information to estimate daily

activities performed by the participant. For that reason, a drop in the accuracy of a predictive model would not be accepted.

We are also interested in other factors, such as the new platform's stability over a period of time. Wearable devices, in common with all other sensors deployed to patient homes, are commonly run virtually 'headless' (that is, without significant visibility to the participant of system status beyond recording status and battery levels). Reliability is therefore as important a key point to us as data validation, so a new version missing more values would not be accepted. The third key element is security. The fourth is the ruggedness of the platform and the extent to which practical ergonomics comply with the requirements of the application domain – for example, is the device adequately comfortable? Is it straightforward to take off and put on? Will it survive if the participant wears it while washing up or taking a shower? Is it adequately straightforward to charge? These factors, however, are largely out of scope for this paper.

B. A planned hardware update

Having deployed wrist-worn sensors in the wild, potential improvements in hardware design were identified. Those improvements are inline with the intented use and purpose of the device. The sensors are expected to provide data that identifies the user's location and wrist movement within a residential environment. Location information is extracted from Received Signal Strength Information (RSSI) collected from receivers scatterred in chosen locations. Wrist movement is extracted by observing the acceleration on the wrist of the user. Micro-electromechanical System (MEMS) accelerometers are monitoring the movement by measuring the acceleration of the wirst in 3-axis.

It is of vital importance for the sensor to generate these data as accurately as possible at a predetermined rate. Thus, considerations has to be made in ensuring the devices operate consistently. There are two fundamental conditions for the device to meet those requirements. Firstly the device has to be worn, and secondly the device has to operate. To meet the first challenge, there is a need to maximise the duration at which the device has to be taken off. Also, users must feel comfortable with having the device on their wrist. From lessons learned from deploying the previous iteration of the sensor [6], devices, unpredictably, had their batteries depleted before their expected battery life. One of the most prominent failure condition is the concurrency of events and the clash of resources usage from different threats on the device. Also, variations between sensors, inherent to manufacturing inaccuracies, introduce further unpredictability.

Based on the motives stated above, the incremental improvements introduced were along the lines of improving the design and reducing unpredictability, while maximizing the battery life of the device. These improvements are well inline with the advancements in battery powerred consumer electronics. Updates in microelectronics design and wireless charging matching were integrated together with improvements in the enclosure design. A rendered version of the all the inhouse developed parts are shown in fig. 1b. The bottom side of the device showing the small footprint of the wireless charging receiver is shown in fig. 1a.

II. Methodology

Given that the device's ergonomic and power characteristics are greatly improved by comparison to the prior version, deployment is considered to be desirable. However, beyond the physical and ergonomic aspects of the device, there are a number of key device characteristics that must be reviewed prior to practical deployment. In particular, under the change management governance procedure applied by the project (which relies on robust release management procedures [7]), it is necessary for us to gain and present evidence that a replacement device is equivalent or better in terms of its performance before we are able to recommend that a proposed replacement be put into active use.

We therefore elect to compare the devices through a small-scale study able to fit within time constraints upon the proposed release schedule. Due to the fact that the wearable sensors are designed for healthcare purpose, the data collection and analysis are expected to be representative of common daily living activities. Therefore, all the data are collected on human rather than physical devices, such as mechanical shakers. During this study, both old and new devices are worn by each participant, and annotated activities are carried out in the order scripted.

The data are then analysed for the following key features: statistical similarity between the datasets, data completeness (i.e. availability of data packets at expected rate and within the expected range) and, perhaps most importantly, the usefulness of the resulting dataset in simple activity recognition tasks making use of the annotated data.

Our change management governance approach requires us to demonstrate that any candidate meet minimum viable standards on all key dimensions before it can be considered as a replacement. That is, since the battery life, ergonomic aspects, etc., are improved with the new device, we would consider it as a replacement in the event that the data reception characteristics and the activity recognition task performance were equivalent. If these prove to be improved, this allows us to make a very strong case for adoption and deployment.

A. Data collection and annotation

In order to collect the necessary dataset, we designed a set of 11 tasks and 4 static wearable positions with a duration between 10 s and 30 s each. Table I shows a description of each task and the amount of time or repetitions. The set of tasks tries to cover a varied set of movements that is representative to common daily living activities. These tasks are selected to reflect the types of task that we expect to come up regularly as proxies for participant health or activity characteristics, such as simple actions and motions that form granular parts of an activity. One task in particular, clapping, is selected to establish wearable sensor behaviour at faster accelerations than achieved by the other actions. In this evaluation, we are not interested in evaluating an exhaustive set of actions, but one in which we could test for discrepancies on the generated data, and in which we could expect. Also, in order to consider

(a) Render and Assembly

(b) Improved Wireless Charging

Fig. 1. SPHERE Wearable 3

human variability factors into the analysis, we asked 4 participants to perform all the tasks.

Each participant completed a full experimental script, with an starting and ending datetime stamp. Inside of an experiment, each activity was also annotated with start and end datetime as well. The time between annotations was fully ignored for the analysis. Figure 2 shows the annotations of one experiment, and the activities that the participant was performing during each time interval.

A second data collection for a long period of time consisted of leaving the wearables sitting on a table for 12 hours in a known, static orientation. This approach has proven to be a useful means to provide a resting noise thumbprint. Additionally, it is a valuable baseline to assess the accuracy of accelerometer readings – in this condition, accelerometer readings should be nearly or entirely constant, so it is a good tool through which to identify either transient fluctuation (noise) or variation with atmospheric conditions, such as temperature. Longer-term testing is also used to explore these issues; we expect to report on this in a later paper.

B. Generated data

Once collected, the following datastreams are available:

- **bt**: date/time of each collected packet of information.
- **ts and tso**: the timestamp and offset provided by the network (their sum corresponds to the bt).
- **accelerometer values**: x, y, z axes in gravity units.
- **rssi**: received signal strength indication as recorded by an environment network present in the house.
- **missing values**: which appear from many potential issues with the system as a result of failure to log values due to instability or from metadata issues causing incorrect labelling.

The synchronisation of both wearables is achieved using a time slotted network (Time Synchronized Channel Hopping (TSCH) Absolute Slot Number (ASN)) and samples are synchronised within 10ms intervals.

C. Activity recognition

The purpose of the experiment is to investigate if both wearables collect enough information to differentiate diverse activities. We will first define some notations, and then introduce the pipeline of the activity recognition. We define each data sequence for activity recognition as a sequence of the 3-dimensional accelerometer data over a time window which is set as 3 seconds with 1 second overlap in this work. Let the number of samples falling in each window be N_s, then each data sequence can be denoted as $x_i \in \mathbb{R}^{N_s \times 3}$, where $i \in \{1,2,\cdots,N\}$. N is the total number of data sequences. The activity recognition is aimed to assign label $y_i \in \mathscr{Y}$ to each x_i. $\mathscr{Y}$ is the label set defined in table I.

Feature Extraction: For both the V.2 and V.3 data, we extract features in each of the accelerometer directions independently on each data sequence. As shown in table II, we define a 12-dimensional feature extractor set, where F_{seg} and F_{sep} belong to frequency domain feature extractors and the remaining are time domain ones. For each accelerometer direction of each data sequence, 12-dimensional features are obtained from the original data sequence. Therefore, after the feature extraction step, the original input data $x_i \in \mathbb{R}^{N_s \times 3}$, $i \in \{1,2,\cdots,N\}$ is replaced by new feature data $z_i \in \mathbb{R}^{36}$.

Classification: With the extracted features $z_i \in \mathbb{R}^{36}$, $i \in \{1,2,\cdots,N\}$, the classification step aims to assign an activity label to each sample. RF algorithm, which is extensively used in a variety of classification and regression problems, is employed as activity recognition classifier in this work. It consists of a large number of individual decision trees that operate as an ensemble [8]. Each individual tree in the RF produces a class prediction and the final prediction takes the class with a majority votes. The Gini impurity, which measures the probability of an incorrect classification given the class distribution, is used as splitting criteria in RF. There are two layers of randomness in RF. Firstly, each individual tree randomly selects samples from the training dataset with replacement, resulting in different trees, which is referred to as 'bootstrap aggregation'. Secondly, when splitting a node in a tree in RF, a random set of features are considered instead of the full feature set, which is referred to as 'feature bagging'. These two-fold randomnesses in RF make the trees individual and uncorrelated, thus leading to the robustness of RF.

For each of the V.2 and V.3 datasets, we split the whole data into 4 partitions in a stratified strategy. Four groups of experiments are conducted respectively on the V.2 and V.3 datasets. In the experiment, each combination of 3 partitions in the 4 partitions acts as training set, and the remaining one is used as testing set. During the classification, 5-folder policy is employed for cross

TABLE I
SET OF ACTIVITIES PERFORMED FOR THE EVALUATION.

Activity short	Description	Time/repet.
table_front	Sitting stationary on table device front	30 s
table_rear	Sitting stationary on table device rear	30 s
table_left	Sitting stationary 'on left side'	30 s
table_right	Sitting stationary 'on right side'	30 s
clapping	Clapping	10 s
waving	Waving	10 s
washing	Washing a cup	30 s
walking	Wearable on wrist, walking	30 s
sit_to_stand	Sit and stand	5×
sitting	Sitting	30 s
stand	Stand still	30 s
kettle	Filling the kettle, then emptying it	5×
high_cupboard	Opening and closing high cupboard	5×
low_cupboard	Opening and closing low cupboard	5×
microwave	Opening and closing microwave	5×

TABLE II
EXTRACTED FEATURES

Designation	Description
F_{min}	Minimum value in the data sequence
F_{max}	Maximum value in the data sequence
F_{sum}	Summation of values in the data sequence
F_{mc}	Mean crossing values in the data sequence
F_{zc}	Zero crossing values in the data sequence
F_{seg}	Spectral energy of the data sequence
F_{int}	Interquartile range of the data sequence
F_{sk}	Skewness of the data sequence
F_{sep}	Spectral entropy of the data sequence
F_{p25}	25th percentile of the data sequence
F_{p75}	75th percentile of the data sequence
F_{kr}	Kurtosis of the data sequence

validation. The final result of each version of wearable is obtained by aggregating the four groups of experiments.

III. RESULTS

A. Data completeness

Figure 3 shows that the number of 'complete' packets per second (that is, packets containing six samples, the transmission protocol used within the SPHERE system) is generally between 4 and 5 per second. This shows a higher number of received packets compared to the previous iteration, which ordinarily sends 4 packets per second. Figure 4 shows that the overall samples per second received on V.3 is also higher than V.2. Bottom fig. 2 shows that wearable V.3 had 1.59 % of missing values on 200 ms windows, while V.2 had 7.23 %. Therefore, there is a notable improvement in the data quality with the V.3 wearable.

B. Platform stability and reliability

Platform stability and reliability are evaluated primarily through a number of longer-term experiments outside the scope of this paper. Ongoing monitoring applications [9] over a period of a fortnight or greater are used to detect early failures or instabilities. However, this small dataset, combined with the twelve-hour test mentioned previously, displays similar findings to those identified in lengthier tests. From fig. 2, we note that 7.32 % of expected data points are missing on the V.2 wearable (labeled C1), whilst 1.59 % of expected points are missing on the V.3 wearable (labeled C0). The V.3 wearable is displaying a significantly lower proportion of transmission or production errors.

A comparison of the two signals shows that the V2 wearable displays a slightly lower peak magnitude in general, possibly due to the presence of a measurement offset or to how the wearables were placed on the arm, as well as indications of slightly lower stability. It can also be identified that the accelerometers are placed along different axes (i.e. the Z-axis is reversed between C0 and C1). Although this should have no direct impact on the quality of machine learning outcomes, it implies that a preprocessing step may be required if the same methods are to be used across both datasets (without retraining).

C. Activity recognition

Figure 5 shows the confusion matrices of activity recognition on wearables V.2 (fig. 5a) and V.3 (fig. 5b) respectively. From the results, we can conclude that both wearables contain enough information for a commonly used machine learning model to predict the activities. The overall classification accuracies over 15 classes are 76.1 % and 79.7 % respectively. It is noteworthy that V.3 wearable achieves 3.8 % higher overall classification accuracy compared with wearable V.2. This might be because the sampling rate of V.3 wearable is 28 Hz, while V.2 wearable is 25 Hz. In addition, from previous analysis, we discovered that V.2 has more sporadic wrong acceleration values, while in V.3 this issue was corrected.

The confusion matrix (fig. 5b) additionally shows that whilst the two wearables perform similarly on most tasks, the V.3 wearable performs better on a particular task – opening and closing a microwave door. This may be a consequence of the lower noise observed on the V3 wearable, since this is a reasonably small and low-magnitude action.

IV. DISCUSSION

The simple method presented here for comparing subsequent generations of bodyworn devices was designed to provide a baseline of evidence on which to base change management decisions within a large project. In the case study given here, we have been able to demonstrate a basis in evidence for accepting a new generation of wearable device as a deployable equivalent for the existing generation, and provide some evidence suggesting that its performance is likely to be better under certain circumstances.

The test itself is short and simple to complete – the data collection itself took approximately half a day in total. The analysis takes somewhat longer, although it is important to recognise that as with any unit test driven approach, the cost of developing the test is likely to be rapidly amortised by the fact that it can then cheaply be reused in subsequent iterations of application. That is to say, the analysis presented here can rapidly be rerun when further changes are made to the system.

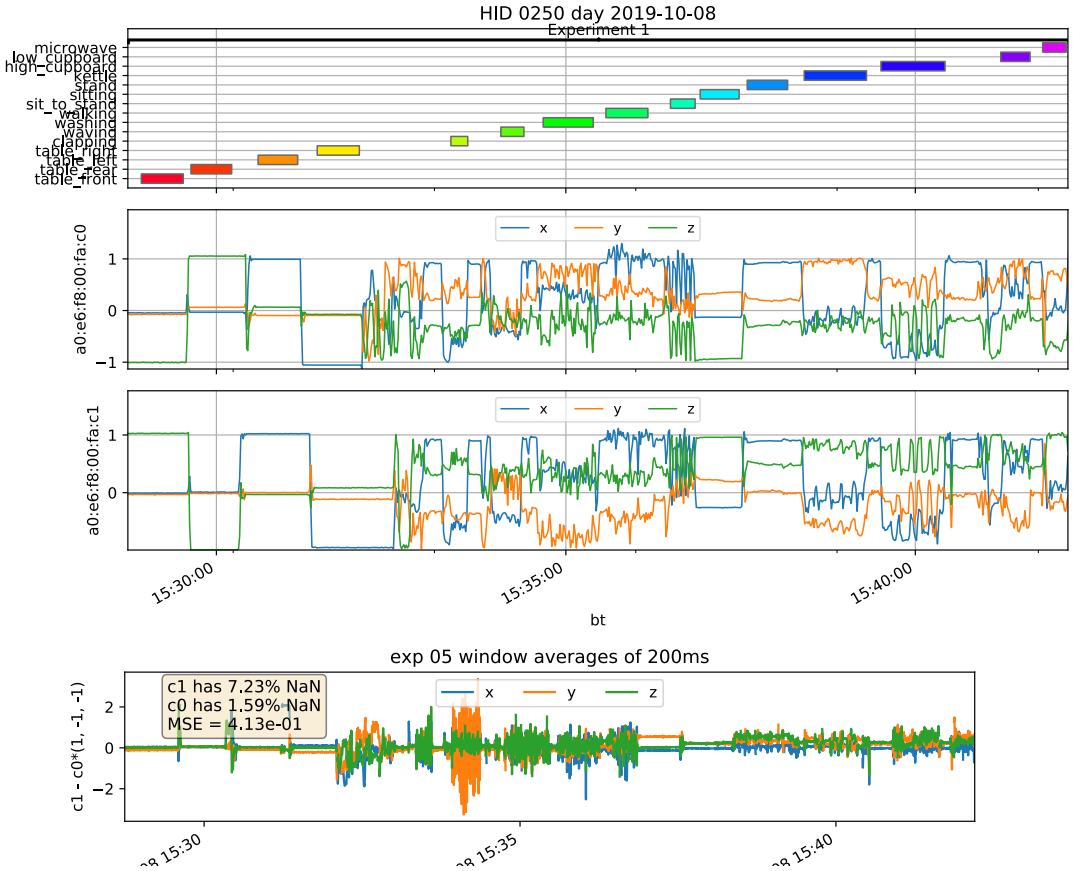


Fig. 2. Annotations of one participant and tri-axial values for two wearables positioned on the right wrist. Upper wearable (MAC addres ending with :C0) corresponds to V.3, while bottom wearable (ending with :C1) corresponds to V.2. It is possible to see that when resting on the table, wearable C0 has the Z axis in the oposite direction than wearable C1. Similarly with Y axis as shown in certain activities like washing the cup, walking, and others.

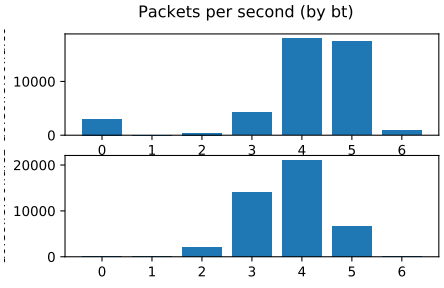


Fig. 3. Number of packets per second (of 6 samples each) received from each wearable.

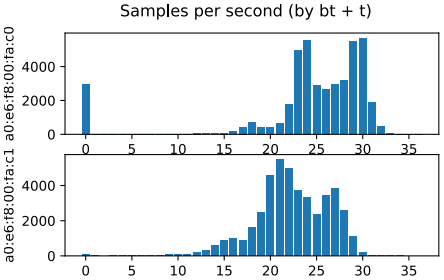
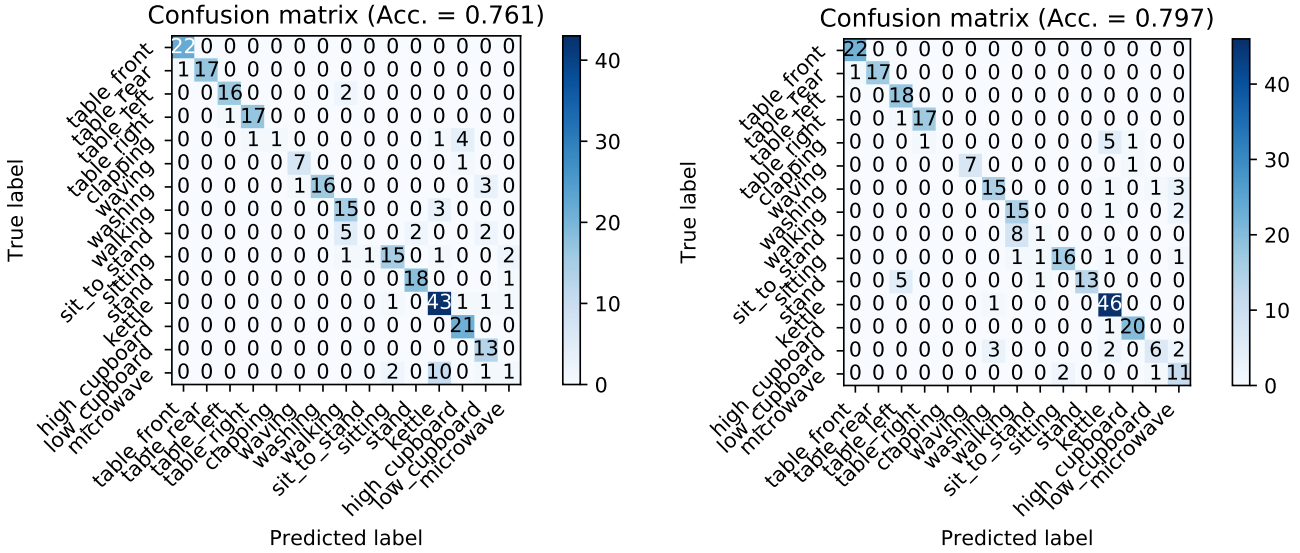


Fig. 4. Number of samples per second received from each wearable.

This process will shortly be repeated to compare the performance of two versions of the V.3 wearable, one of which is equipped with a coil for wireless charging. In this manner we will evaluate the extent to which the addition of a charging coil and ferrous pad alters the performance of the device. We also expect to make use of this same methodology with the addition of location annotation to compare the RSSI localisation performance of the three different classes of device (V.2, V.3, V.3 with charging pad), a separate issue that was not evaluated in this test methodology.

A key question, however, that remains extant is whether the evidence provided by this testing approach is adequate, or whether further shortcomings of the data or the device form factor may arise that have not yet been tested for. This is a common problem in software testing [10]. Our solution to it, as hinted at by Zhu et al, is to regularly review real-world device performance and issues as they arise and use them as a resource from which to revise our requirements and specifications. We then return to our testing methodology and review the metrics and activities under test, to ensure that any points that we have not covered thus far will be added to our testing methodology in the future.



(a) V.2 wearable

(b) V.3 wearable

Fig. 5. Confusion matrix for the different versions of the wearable device.

V. Conclusion

An evidence-based approach to acceptance testing provides an increased confidence in the hardware provided, as well as an early opportunity for machine learners to work actively with the dataset returned by each wearable. This has proven to be a good opportunity to identify and address bugs in device firmware or software at an early stage, as well as a chance for knowledge transfer between teams. Systematisation and democratisation of testing between teams has allowed us to improve our understanding of system requirements and address potential sources of error down the line. Finally, this methodology provides us with a blueprint for approaching similar acceptance testing tasks into the future.

Acknowledgment

This work was performed under the SPHERE IRC funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/K031910/1 & under the Elizabeth Blackwell Institute for Health Research, University of Bristol and the Wellcome Trust Institutional Strategic Support Fund, Grant 204813/Z/16/Z.

References

[1] Xu Ye, Guanling Chen, and Yu Cao, "Automatic eating detection using head-mount and wrist-worn accelerometers," in *2015 17th International Conference on E-health Networking, Application Services (HealthCom)*, Oct 2015, pp. 578–581.

[2] D. Kelly, S. McLoone, B. Logan, and T. Dishongh, "Single access point localisation for wearable wireless sensors," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2008, pp. 4443–4446.

[3] A. Pantelopoulos and N. Bourbakis, "A survey on wearable biosensor systems for health monitoring," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2008, pp. 4887–4890.

[4] G. Andreoni, A. Fanelli, I. Witkowska, P. Perego, M. Fusca, M. Mazzola, and M. G. Signorini, "Sensor validation for wearable monitoring system in ambulatory monitoring: Application to textile electrodes," in *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare*, ser. PervasiveHealth '13. ICST, Brussels, Belgium: ICST, 2013, pp. 169–175.

[5] K. L. Wong and D. L. Lindstrom, "Off the bathtub onto the roller-coaster curve (electronic equipment failure)," in *1988. Proceedings., Annual Reliability and Maintainability Symposium,*. IEEE, 1988, pp. 356–363.

[6] X. Fafoutis, A. Elsts, A. Vafeas, G. Oikonomou, and R. Piechocki, "On predicting the battery lifetime of iot devices: experiences from the sphere deployments," in *Proceedings of the 7th International Workshop on Real-World Embedded Wireless Systems and Networks*. ACM, 2018, pp. 7–12.

[7] M. Sallé, "It service management and it governance: review, comparative analysis and their impact on utility computing," *Hewlett-Packard Company*, pp. 8–17, 2004.

[8] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] N. M. Vichare and M. G. Pecht, "Prognostics and health management of electronics," *IEEE transactions on components and packaging technologies*, vol. 29, no. 1, pp. 222–229, 2006.

[10] H. Zhu, P. A. Hall, and J. H. May, "Software unit test coverage and adequacy," *Acm computing surveys (csur)*, vol. 29, no. 4, pp. 366–427, 1997.