Chapter 16
User-Centred Evaluation for Machine Learning

Scott Allen Cambo and Darren Gergle

Abstract Activity tracking wearables like Fitbit or mobile applications like Moves have seen immense growth in recent years. However, users often experience errors that occur in unexpected and inconsistent ways making it difficult for them to find a workaround and ultimately leading them to abandon the system. This is not too surprising given that intelligent systems typically design the modelling algorithm independent of the overall user experience. Furthermore, the user experience often takes a seamless design approach which hides nuanced aspects of the model leaving only the model's prediction for the user to see. This prediction is presented optimistically meaning that the user is expected to assume that it is correct. To better align the design of the user experience with the development of the underlying algorithms we propose a validation pipeline based on user-centred design principles and usability standards for use in model optimisation, selection and validation. Specifically, we show how available user experience research can highlight the need for new evaluation criteria for models of activity and we demonstrate the use of a user-centred validation pipeline to select a modelling approach which best addresses the user experience as a whole.

16.1 Introduction

Activity tracking systems such as wearable devices like Fitbit and Jawbone or mobile applications like Moves and Google Fit have seen extraordinary growth in commercial activity over the past several years. Yet, a common problem with these systems is early user abandonment – shortly after initial adoption many users stop

S. A. Cambo (✉) · D. Gergle
Northwestern University, 2240 Campus Drive, Evanston, IL 60208, USA
e-mail: cambo@u.northwestern.edu

D. Gergle
e-mail: dgergle@northwestern.edu

J. Zhou and F. Chen (eds.), *Human and Machine Learning*, Human–Computer
Interaction Series, https://doi.org/10.1007/978-3-319-90403-0_16

engaging with the tracked information or even stop using the system entirely. To better understand why this is happening, researchers have begun to survey users of activity trackers [10, 27], conduct detailed user interviews [23], and synthesise feedback from online product reviews [27]. Across these studies, users often report that inaccuracy and errors play a big role in their decision to abandon their activity trackers [10, 23, 27] and that they are uncomfortable with activity tracking performance, system behaviour and overall user experience.

To address these challenges, user experience researchers have recommended integrating interactive elements into activity tracking systems that permit the users to better understand errors and play a more active role in calibrating the system. Some health tracking systems researchers have gone a step further to argue that manual calibration could actually aid the user to better understand their tracked and inferred data [7, 10, 23, 27]. However, there is a gap in our understanding that exists between the design of these interactive aspects of the user-experience and the performance of underlying algorithms and models of activity tracking systems. In particular, computational techniques such as model personalisation and active learning—which are inherently suited to the integration of user interaction—tend to be developed and evaluated with a focus on model accuracy instead of considering the broader implications of performance and how it relates to the user experience and interaction with the activity tracking system.

In this chapter, we show how a user-centred approach to the evaluation of model personalisation techniques could help bridge the gap that exists between the way we research and develop the user interaction and the way we research and develop the underlying model. User-centred evaluations like those we describe in this chapter can lead to designs and technical implementations that better enable users to have a richer and more fulfilling interactive experience with their activity tracking systems. In Sect. 16.2, we contrast the technical performance perspective of model evaluation with the user experience perspective derived from research on why people abandon activity trackers. We use research in health technology design and health behavior change to motivate the need to identify *seamful design* opportunities in the system's underlying algorithms. In Sect. 16.3, we use these principles to define a validation algorithm that provides an individualised view of model performance and demonstrate how it can be used in model optimisation, validation, and selection. Then we identify seamful design presentations of model confidence which can help address the user challenges described in Sect. 16.2. Finally, in Sect. 16.4, we describe how these principles of usability along with the user-centred validation process help us make model selection decisions that address the whole user experience and not just the model validation. Throughout this chapter we discuss approaches that may support better visibility by making model behaviour more salient to the user and better transparency by aiding the user's understanding of model behavior. We further see visibility and transparency as prerequisites for users to gain the trust needed to use systems for sustained behavioural change.

16.2 Background

16.2.1 A Technical Performance Perspective on Evaluation

A common evaluation approach used in much of the technical literature examining activity tracking systems is to treat model performance as a binary construct (i.e., correct or incorrect) and to minimise errors at the aggregate level in an effort to optimise activity tracking performance for the user base as a whole. Consider the case of a wearable device used to infer activity from a stream of motion sensor (accelerometer) data. In building such a system, researchers collect examples of accelerometer data associated with the activities they are interested in and then build a model using a supervised machine learning technique. In order to measure how well the resulting model recognises the correct activity from the accelerometer data, each incorrect prediction is recorded as an error. Researchers then assess model performance by examining the overall error rate or its inverse, accuracy. Using this measure, they can decide between competing modelling approaches and determine which will work best for their given activity recognition application. Subsequent iterations and refinements of these algorithms may be optimised based on this same measure – but a question exists as to whether this binary and aggregate view of errors is descriptive enough when considering the entire user experience.

16.2.2 A User Experience Perspective on Evaluation

While many technically oriented approaches to errors in the context of activity recognition focus on minimising errors and optimising accuracy, it's important to keep in mind that the larger goal of these systems is often to help users track their behavior in such a way that patterns will emerge to help them make better decisions concerning their activities and health related behaviors. One way to focus model evaluation on this more user-centred goal is to go beyond the simple binary treatment of errors and consider more graded measurements such as prediction likelihoods for individual users and corresponding confidence measures. These metrics can serve a dual purpose which is to provide more detail about model performance and to provide more clarity to the users about how or why a given model prediction may be off.

It's also important to go beyond a single aggregate measure of performance and think carefully about the distribution of performance scores and how a given model affects individuals or groups of users. As an example, consider two models with similar accuracies - the first with an overall accuracy of 80% and the second with an overall accuracy of 70%. The first model may seem the easy choice based on aggregate performance. However, if the distribution in the first model is unimodal and falls within a narrow range (e.g., [77–83%]), and the second model is bi-modal with two narrow ranges (e.g., [43–47%] and [93–97%]), the decision becomes much more complex. Other aspects of the performance distribution may also affect what is considered "the best" model depending on the user goals. Should researchers pick the model that does the best (on average) for everyone or one that ensures the worst performance for an individual is still above a certain threshold? The correct answer to such questions requires looking closely at the distribution and likely depends on the particular application and the user's intended goals and tasks. Once the technology moves beyond proving that an algorithm is capable of learning the target concept to proving that it can help humans improve their health, errors become more complex than a simple binary construct of correct or incorrect and optimisation isn't as simple as choosing the best aggregate performance measure.

User Expectations and Perceptions of Errors

In addition to the ways in which we measure errors and consider model performance, it's also important to consider users' expectations and their perception of errors when using activity tracking systems. Users begin forming their expectations from activity tracking systems with marketing and advertisement materials such as Fitbit's slogan "Every Beat Counts" and as a result they often initially expect near perfect accuracy. As soon as users begin interacting with their activity trackers, they begin developing a mental model of the underlying algorithm formed by both correct predictions and errors [10, 23, 27]. These errors do not immediately cause users to abandon their activity trackers; instead, users take this to be a limit of the technology's capabilities and consider new behaviours or workarounds, which allow them to continue getting value from the activity tracker. One example of a workaround comes from a participant in Harrison et al.'s interview study [10], *"I was trying to figure out where I could put my Fitbit to get what I thought was a good amount of steps. I was yknow, putting it in my sock, putting it in my pants pocket, tying it to the cuff of my pants […] I was also pedalling backwards whilst going down hills"*.

Users also test their understanding of the system by trying to replicate correct and incorrect predictions which should align with their understanding of the boundaries of the model [23, 27]. When the performance of the system continues to violate expectations that it will perform consistently, as advertised, or as designed, users begin disengaging from the system until they eventually abandon it [27]. Another way to think of this is that each violation of expectation erodes the trustworthiness of the model. As previously alluded to, one way to help the users set better expectations is to provide them with richer details regarding the predictions and model confidence – a design decision that involves consideration of technical components such as model selection. These details can provide greater model transparency that help the user to understand what they can and cannot expect from their activity trackers.

User expectations and experiences are also highly variable. Variations in individual lifestyle, physical characteristics and values for the tracked information can lead users to different conclusions about whether the accuracy of the activity tracker is adequate for them to derive meaningful use from it. A common conclusion users make when they encounter errors is that the system was designed for someone else. This point is well captured by an interviewee from Yang et al.'s study [27]: *"The [FitBit] Flex does not count steps if you are pushing a stroller or cart. This may not* *be an issue to some, but it is to me b/c I have stroller age children. Then I noticed it was logging hundreds of steps while I was rocking my baby. Maybe something else out there will be more accurate for my lifestyle."* In such cases, the ability of a model to learn personal idiosyncrasies or adequately present the likelihood of being correct or incorrect is likely to be an important evaluation criteria.

The Burden of Interaction or the Benefit of Engagement?

Up until this point in the chapter, we have mainly discussed the challenges that errors present. However, some researchers argue that engagement with errors may actually be beneficial for the end-users of activity tracking systems [7]. An original vision of many ubiquitous computing technologies, such as activity tracking, focused on ways to make the technology fade into the background, helping the user to increase focus on the task at hand and decrease focus on the tool designed to help with that task. To achieve this, research has pushed the boundary of what we can infer automatically about user context from passively collected sensor data using machine learning and artificial intelligence. In the initial vision of ubiquitous computing environments, this is referred to as a *seamless design* approach in which particular technical nuances of a tool (or model) are made invisible to the user [25]. In [17], we see this design philosophy in action as the authors recommend a fully-automated approach to the personalisation of an activity recognition model (semi-supervised learning) over a semi-automated and interactive approach (active learning), even though model performances were comparable stating that it would be too burdensome for the user to interact with such a system.

Other researchers [10, 23, 27] have argued that the way in which users fail to develop a working understanding of the underlying algorithms, and the way in which users attempt to find workarounds that make the system more useful, imply that the design of activity tracking systems should make the underlying model more visible and include more interactive features that allow the user to calibrate the system to their personal idiosyncrasies and requirements. Choe et al. [7] state that fully automated tracking is not only difficult (or impossible) to achieve in some scenarios, it ignores the potential for synergies between the user's goal (understand and change health related behaviour) and the goal of the activity recognition model (accurately predict the appropriate activity label from the user's smartphone motion sensors). In contrast to seamless design approaches, *seamful design* has been proposed as a design philosophy which aims to strategically build "seams" into the user interface. These seams represent aspects of the model that the user can utilise to get a better sense of the underlying algorithms. Activity trackers employing seamless design can fail silently such that the user has no way of knowing that the knowledge generated by the system is no longer valid until it becomes evident when the output displayed is outside the boundaries of the user's expectations. While seamful design complicates the design and interaction process, it has the potential to make errors more visible making the limitations of the system salient such that corrections can be easily made to the tracked information. These additional opportunities for awareness and engagement are likely to have the positive side effect of continued use and lasting health change as observed by [19].

Seamful design of intelligent systems such as activity trackers requires a user-centred approach to *align the underlying mechanics* used to derive the "intelligence" of the system *with the user interface and experience*. In the next section, we'll discuss different algorithms that have been proposed for future activity tracking systems, how they are likely to affect the overall user experience and how we can better validate them with respect to the user experience before implementing the entire activity tracking system.

16.2.3 How Model Selection Affects the User Experience

To understand how seamful design may be achieved in activity recognition technologies, we discuss how the user experience research described in the previous section can help guide us in the research and development process at the stage of algorithm development and model selection. To help structure how to do this, we consider various facets of common usability and user-centred design principles in the context of model development and algorithm selection. Then, we describe the potential modelling approaches in terms of what they afford the design of the user interface and user experience.

User-Centred Design for Machine Learning

One of the goals of this chapter is to motivate and highlight a user-centred design process for developing ML-driven technology in which there is more of a focus on the final user experience at the earliest stages of research and development. We draw inspiration from the following principles of user-centred design as specified in the International Usability Standard, ISO 13407 and consider them in the context of model selection and algorithm design for activity tracking:

1. The design is based upon an explicit understanding of users, task and environments.
2. Users are involved throughout the design and development.
3. The design is driven and refined by user-centred evaluation.
4. The process is iterative.
5. The design addresses the whole user experience.
6. The design team includes multidisciplinary skills and perspectives.

These principles suggest a process by which we iteratively incorporate knowledge of and by the user into the design of the underlying learning algorithms and models. The first principle encourages an initiative to understand the user's relationship to the technology by answering questions like: *"Why will someone use this technology?"*, *"What will they use it for?"* and *"In what context will they use it?"*. For activity recognition, we can begin to answer these questions by drawing on research into how people use and abandon activity tracking systems as well as research on why tracking desirable and undesirable behavior is expected to have a positive impact on the user's health. The second principle aims to directly include feedback from the user during all stages of design and development. Doing this at the algorithm development and model selection stage is challenging, because these components are generally decoupled from the user interface. However, we can use knowledge of users, their tasks and their environment to develop user stories which can help direct the development of algorithm evaluation metrics such that they better reflect the expected model performance and user experience.[1] For example, one of the users we quoted in the previous section tells a story about how she feels that the activity tracker was not intended for someone like her because it demonstrated that it could not recognize the steps she took while pushing a baby stroller. This story should indicate to the algorithm developer that the model has likely learned to distinguish activities based on arm movement and that this arm movement is not always present in cases of everyday walking. By looking at model accuracy from the perspective of accuracy for an individual, we can begin to see how the model might address this aspect of the user experience.

16.3 Experiments in User-Centred Machine Learning for Activity Recognition

In this section, we describe how we can apply user-centred design principles to the algorithm design and model selection stage of development even when this stage is decoupled from the end user. Here, we demonstrate two experiments which were designed to help us refine the modelling approach from a more user-centred perspective. For each experiment, we first describe the expected effect each modelling approach will have on the user experience and what potential there is for seamful design. Then we define an evaluation algorithm which allows us to compare modelling approaches with respect to the expected individual model performance and user experience. To reflect the current research in activity recognition, which proposes individual models of activity rather than a single general impersonal model for all, the evaluation algorithms we define are designed so that we can analyse distributions of model performance as opposed to aggregate measures.

[1] In user-centred design, the term "user stories" refers to a set of scenarios (sometimes fictional) that best reflect common experiences of the target user in the context of their task and environment.

16.3.1 Experiment 1: Impersonal, Personal and Hybrid Models

For an activity tracking system to understand when a user is performing a particular activity, it must have some model of how the sensor data relates to that activity. Typically this model is constructed in a supervised approach where sensor data is collected in a controlled setting so that the ground truth label can be easily observed and recorded. The motivation for this approach is that by collecting data in the lab for many different people, the model might learn characteristics of the relationship between the sensor data and the activity that can be applied generally to data from new users who were not observed in the initial data collection process. In activity recognition, this is referred to as an *impersonal model*.

In contrast, *personal models* use only training data representing an individual end-user obtained through manual interaction. While personal models often perform with better accuracy than impersonal models, researchers are often reluctant to recommend the approach since the labelling task could be considered burdensome by the user [4]. Furthermore, the models can be brittle if the user hasn't labelled activity cases in varying contexts or environments. Personalised or *hybrid models* (sometimes called mixed models) have been proposed as a compromise in which the system is initially deployed with an impersonal model that becomes increasingly personalised by incorporating observations of new data that better represent the end-user.

The *impersonal*, *personal* and *hybrid* modelling approaches each present different possibilities for the user interaction design. Since impersonal models are the current commercial standard, use by consumers of the commercial product can be studied to understand how impersonal models affect the user experience [10, 23, 27]. From this research, we can expect that users are reasoning about errors in ways that make it difficult for them to apply the information from activity tracking systems. Increasing the sophistication of the learning algorithm may increase prediction accuracy from a traditional validation perspective, but this does not necessarily result in a better understanding of personal health behaviour that can be used to make better health decisions.[2] Alternatively, most personal and hybrid modelling approaches require that the user manually label recorded activity data. This could take place in an initial calibration phase or as an ongoing and dynamic process integrated with user interface features that are designed for the continued engagement and awareness that leads to better health outcomes as described by [7]. In experiment 1, we aim to recreate the experiment described in [15] and extend the analysis to achieve the following:

[2] There are also reasons to believe that there may be a ceiling to the accuracy of impersonal model performance. Yang et al. suggest that one barrier to better impersonal model accuracy is inherent in the translation of medical grade tracking equipment to the consumer market which prioritises ergonomics, size, fashionability and many other factors over accuracy [27]. Lockhart et al. suggest that as the number of individuals represented in the impersonal training dataset approaches 200, the increased accuracy gained from each individual decreases and plateaus around 85% accuracy [15].

1. Understand the distribution of expected model performance for individuals represented in the dataset instead of a single aggregate measure of model performance.
2. Compare the expected benefits (increased model performance) and expected burden (increased user interaction) exhibited by either a personal or hybrid model.

A User-Centred Validation Pipeline for Experiment 1

Algorithm 1 User-Centred Validation for Experiment 1

```
Let $D$ be all the labelled activity recognition data
    we have available to study
$D_{personal}$ will represent the subset of data representing a user, $u$

for all $u$ in the set of users in $D$
    Let $D_{personal}$ be all data in $D$ where $D_{personal} == u$
    Let $D_{impersonal}$ be all data in $D$ where $D_{personal} != u$

    Let $T_{personal}$ be the subset of $D_{personal}$
        which was sampled through some sampling function,
        $s(D_{personal})$ for training the personal model

    Let $T_{impersonal}$ = $D_{impersonal}$ to use all
        available impersonal data

    Let $T_{hybrid}$ be the training set which combines data from both
        $T_{personal}$ and $T_{impersonal}$ by joining the sets
        through some function $j(personal, impersonal)$

    Let $V_{personal}$ be the subset of $D_{personal}$
        which was sampled through some sampling function,
        $s(D_{personal})$ for testing or validating all models

    Let $\theta_{personal}$ be the model trained on $T_{personal}$
    Let $\theta_{impersonal}$ be the model trained on $T_{impersonal}$
    Let $\theta_{hybrid}$ be the model trained on $T_{hybrid}$

    Make predictions on $V_{personal}$ using $\theta_{personal}$ and record for analysis.
    Make predictions on $V_{personal}$ using $\theta_{impersonal}$ and record for analysis.
    Make predictions on $V_{personal}$ using $\theta_{hybrid}$ and record for analysis.
```

Algorithm 1 begins by iterating through a set of validation users. This set of validation users can be the same as the training set as long as the user is held out as would be done in a leave-one-user-out validation process. For each iteration, we separate the personal data, $D_{personal}$, from the impersonal data, $D_{impersonal}$. $D_{personal}$ is then sampled through some sampling function, $s(D_{personal})$, to create independent training, $T_{personal}$ and validation, $V_{personal}$ datasets. To put a user-centred perspective on this, $T_{personal}$ is attempting to represent the data that the user might have labelled during an initial calibration phase while T_{hybrid} is attempting to represent the data

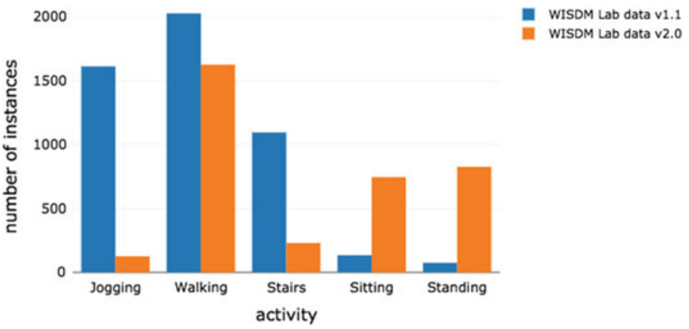

Fig. 16.1 This figure represents the number of labels of each activity class we have in each WISDM dataset

that the user might have labelled during an ongoing data collection interaction. The sampling function, $s(D_{personal})$, should be a representation of the process by which the user discloses the label for a particular activity. A first validation approach might use a random sampling function to approximate which personal training instances get disclosed. Using the validation pipeline with enough iterations to generate stable metrics for a stochastic process like random sampling, we can start to approximate how the model building process affects performance for various individuals.

To demonstrate the utility of our pipeline we use it to assess the importance of hybrid and personalised models as described by Lockhart and Weiss [15]. Specifically, we want to evaluate Lockhart and Weiss's idea that a model, $\theta_{personal}$ where the amount of personal data, $T_{personal}$, is much smaller than $T_{impersonal}$, is preferable to both, $\theta_{impersonal}$, and, θ_{hybrid}, when the sampling function, $s(D_{personal})$, is a function that samples at random and tries to preserve the overall class distribution in $D_{personal}$. For this part of the experiment, we iterate on the sampling of the personal data, training of the models and testing on the validation sample four times and report the mean accuracy for each user (Figs 16.3, 16.4 and 16.5).

The Data and Modelling Approach

We use publicly available datasets by WISDM lab (Wireless Sensor Data Mining) to perform our analysis. The first dataset, v1.1, represents a study in which data was labelled by asking participants to perform various activities with a smartphone in their front pocket recording all movement with a tri-axial accelerometer sampling at 20 hertz [12]. In this study, the data were labelled by a research assistant while the participant performed the activity. The second dataset, v2.0, represents a study in which a modestly interactive system called ActiTracker was developed to allow users to label their activity data in the wild using pre-determined activity labels [16, 26] in order to create personalised activity recognition models. Fig. 16.1 shows the distribution of labels for each class by dataset. WISDM v1.1 includes 35 users who have labelled more than one activity type while v2.0 includes 25. For more information on the WISDM datasets we used refer to [15, 16, 26].

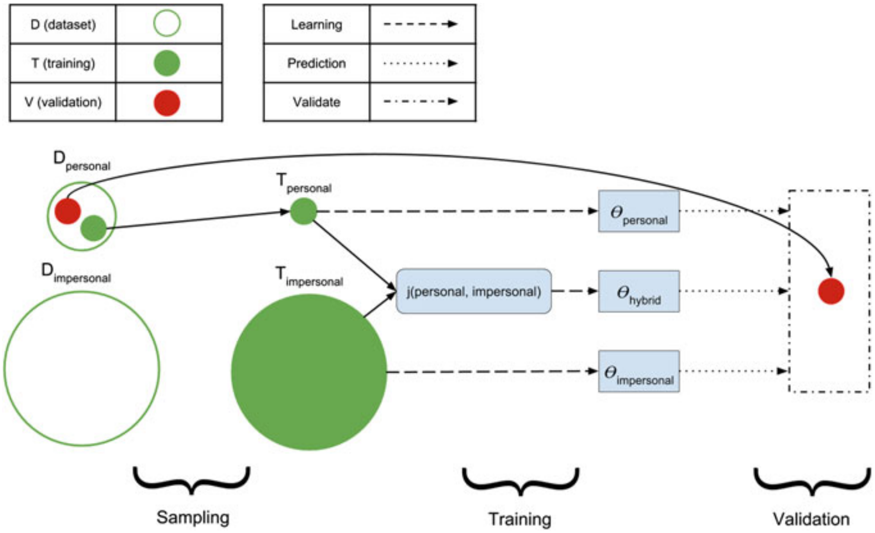

Fig. 16.2 This figure visualises the sampling, training and validation process

In the development of activity tracker systems, we need to consider how developers should go about curating the data for the impersonal dataset (Figs. 16.1 and 16.2). Many context-aware systems will do this by paying participants to perform the activities in a setting where the ground truth activity can be easily labelled with the intention of using the model trained on the data from an initial set of paid users to provide a working model for the first set of unpaid users. We reflect this in these experiments by using the WISDM v1.1 dataset as the impersonal dataset while iterating through v2.0 as the personal dataset. While both datasets included the activities “jogging”, “walking”, “sitting”, “standing” and “stairs”, the v1.1 dataset differentiated “stairs” into “up stairs” and “down stairs”, while the v2.0 dataset included a “lying down” activity label. To resolve these differences, we removed the instances labelled with “lying down” from the v2.0 dataset and consolidated the “up stairs” and “down stairs” classes into a “stairs” class similar to the first dataset. The final class distribution for each dataset can be seen in Fig. 16.1.

To best replicate the work of Lockhart and Weiss, we used the Random Forest classifier as implemented in the Scikit Learn module for machine learning in Python [20]. For all the experiments presented in this paper, we use the parameters described in [15] unless stated otherwise. However, Random Forest models randomly sample the training instances to create simple decision trees (shallower depth, fewer features to consider at each split) and then average the results to maximise predictive accuracy while mitigating over-fitting. This means that the predictions can be inconsistent with the exact same input data. To ensure more consistent results we use 1000 decision trees or more.

In assessing accuracy we take our modelling process to be a multi-class classification task in which each label is assumed to be mutually exclusive. To validate the expected prediction accuracy for each user, we simply take the number of prediction errors, e, from a model on a user's validation sample, $V_{personal}$, subtract it from the size of the validation sample $|V_{personal}|$, and finally normalize the result by the number of validation samples.

$$accuracy = \frac{|V_{personal}| - e}{|V_{personal}|}$$

Since the sampling for both $V_{personal}$ and $T_{personal}$ is randomly sampled with replacement, we repeat the modelling process 10 times for each user and report the mean.

Experiment 1 Results and Discussion

Figures 16.3, 16.4 and 16.5 are box plots that represent the distribution of model accuracies among different users on the y-axis and the number of personal training samples along the x-axis. The red, blue and green boxes represent the impersonal, personal and hybrid models respectively with dots representing mean accuracy across random sampling iterations for each user. Similar to Lockhart and Weiss, Fig. 16.3 shows that the impersonal model has the lowest accuracy in nearly all scenarios. A closer look on an individual basis reveals that the user receiving the best performance is at 79% accuracy, while the modal user performance is 45% and the worst individual performance is at 3%. The hybrid and personal models each considerably outperform the impersonal model even with only 5 samples from the end-user's labelled data

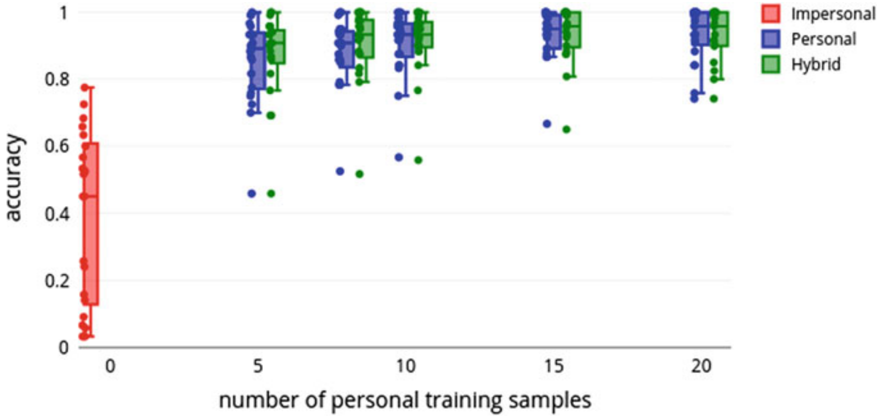

Fig. 16.3 These box plots represent the distribution of accuracy measurements across participants given a Random Forest Classifier, when trained with an impersonal, personal and hybrid dataset

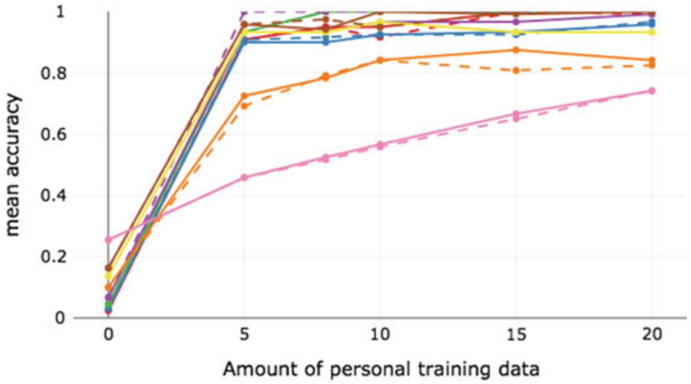

Fig. 16.4 Similar to Fig. 16.3 this plot highlights those whose expected impersonal accuracies were lowest and shows how this accuracy progresses as we increase the number of personal training samples for a particular user. The solid lines represents the personal model and the dashed lines represents the hybrid model. Each color represents a different user

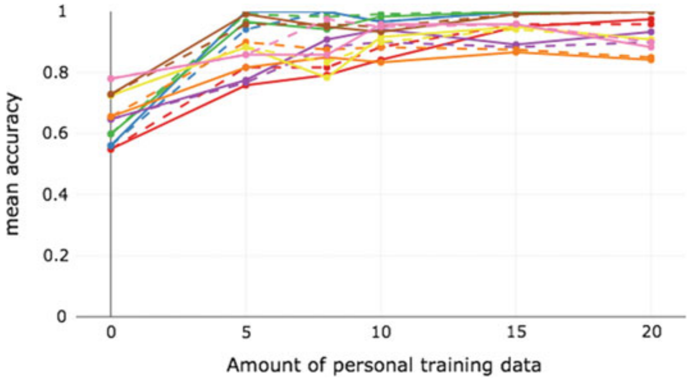

Fig. 16.5 Similar to Fig. 16.4 but with those whose expected impersonal accuracies were highest. The solid lines represents the personal model and the dashed lines represents the hybrid model. Each color represents a different user

with which to use for training. With each sample being equivalent to 10 s of activity being labelled, 5 samples equates to 50 s of labelled activity.

In comparing the personal and hybrid models, we see some differences when the model is trained with 5 or 8 personal samples with a higher inner quartile range for the hybrid model in both cases and a tighter inner quartile range (85–95%) for the hybrid model with only 5 training samples. However, what differences there might be between the personal and hybrid models appear to go away as the number of samples increases beyond 10. This translates to approximately 100 s of user labelling – which may or may not be burdensome for the user depending on the overall fitness level and demographic characteristics of the users. For example, a minute and a half of jogging for a relatively fit undergraduate student may not be very burdensome at all. However, the same amount of required labelling activity may be excessively burdensome for an older adult that is not used to exercising and is recovering from knee surgery.

The distribution *within* a modelling technique is another aspect of model performance that should be explored. By individually evaluating the expected model performance for users, we can see that there seem to be two clusters concerning impersonal model accuracy with one achieving higher than 40% accuracy and the other achieving lower than 25% accuracy. To understand how individuals in each of these groups may benefit from either a personal or hybrid model, we highlight the eight individuals who get the worst expected performance from the impersonal model in Fig. 16.4 and the eight individuals who get the best performance from the impersonal model in Fig. 16.5. In Fig. 16.4 we can see that seven out of these eight participants would have found that the model is nearly always incorrect. We also see that whether we incorporate impersonal data or not (i.e., creating a hybrid or personal model), the accuracy of a model quickly exceeds 90% for all but two users with only 5 labelled instances from the user making the labelling indispensable for those who would have had terrible performance using the impersonal model. In Fig. 16.5, we find that users who were getting better performance with the impersonal model will likely get less of an increase in performance as they label their activity data.

Often when we discuss the potential benefit of hybrid models, it is in the context of reducing the burden of labelling. In our experiment, we do not observe evidence that hybrid models substantially reduce the labelling effort presented by personal models which would present itself in Figs. 16.4 and 16.5 as a datapoint higher than its counterpart of the same colour on the dashed line.

To quantify the concepts of "burden" and "benefit" for the purpose of algorithm optimisation, we might consider "benefit" to be the increase in model performance and "burden" to be the amount of additional interaction or labelled instances required. This "benefit-to-burden" ratio for a particular personal or hybrid model should help us make decisions at the level of algorithm design with a user-centred perspective. It is important to note that "burden" and "benefit" as we define them here are not fully representative of how these concepts play out in the user experience. As we mentioned in Sect. 16.2.2, these interactions can help the user achieve the goal of changing their health related behaviour. Rather we make this naïve reduction in order to present a tractable algorithm optimisation metric to aid the algorithm design process. To fully understand the dynamic between burdensome and beneficial interactions in intelligent systems, additional user experience research is required with working prototypes in the context of a final user interface and interaction design. Furthermore, richer aspects of burden (e.g., those that account for physical exertion as it relates to the intended user population and demographics) could be integrated into a notion or measure of burden in a way that better matches the user experience goals of the system.

16.3.2 Experiment 2: Model Confidence for Seamful Design and Model Personalisation

Seamful design, as described in Sect. 16.2, requires that appropriate affordances of the underlying algorithms and models are identified as a potential "seam". This means that there exists the possibility for a user interface design which exposes this aspect of the model such that it can be appropriated by the user for their task without the need for expert technical knowledge of the system. In experiment 1, we see that a small amount of labelled personal data yields a great improvement in model performance compared to impersonal models when these labels are chosen at random. However, users are likely not choosing which moments to label at random. From [27] we see that users typically begin thinking about interactions which will improve the model performance when they observe errors in the system. In experiment 2, we aim to demonstrate a way in which model confidence, the model's ability to assess the likelihood that its prediction is correct, can help guide both the user's understanding of the system and the user's manual tracking behaviour. Specifically, we design experiment 2 with the following objectives:

1. To understand the potential for model confidence to be used in the seamful design of activity tracking.
2. To understand the potential for model confidence to aid model personalisation through either an active learning or semi-supervised learning approach.

In integrating the concept of seamful design to algorithm evaluation, we can draw insight from work by Chalmers et al. that has explored seamful design in the context of presentation approaches to model certainty along the following facets [5, 6]:

- *optimistic*: show everything as if it were correct.
- *pessimistic*: show everything that is known to be correct.
- *cautious*: explicitly present uncertainty.
- *opportunistic*: exploit uncertainty.

Currently available activity trackers use an *optimistic* approach to the presentation of model confidence. *Pessimistic* and *cautious* approaches can help users understand the limitations and strengths of the model. Additionally, these approaches can give the user a sense of moments when they should rely on their own accounting of their activity instead of the system's.

Opportunistic presentation of model confidence lends itself to a research topic within machine learning called *active learning* in which the system attempts to select unlabelled observations that would be most beneficial to model performance if they were to be labelled. We can think of this as selecting the samples which present the best benefit-to-burden ratio. One of the most common ways of deciding which sensor observations should be labelled is *least confident sampling*. In theory, least confident sampling helps the user to know which instances increase model performance the most by informing the user when the likelihood of predicting a sensor observation correctly is low. This is done by notifying the user of model

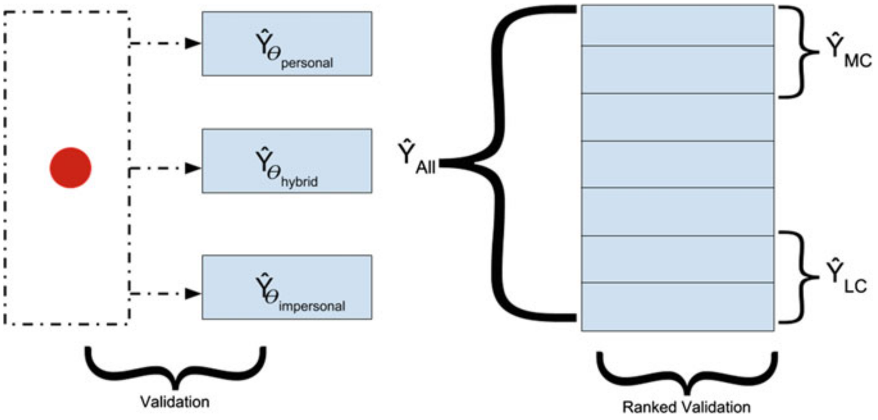
$\hat{Y}_{\theta_{personal}}$
$\hat{Y}_{\theta_{hybrid}}$
$\hat{Y}_{\theta_{impersonal}}$
$\hat{Y}_{All}$
$\hat{Y}_{MC}$
$\hat{Y}_{LC}$
Validation
Ranked Validation

Fig. 16.6 This figure shows how the confidence ranked predictions are partitioned in order to best assess the model's ability to be accurate when considering its most confident (MC) and least confident (LC) predictions

confusion and a labelling opportunity. One point of caution in using least confident sampling is that a model that does not represent the target concept (activity) well can have the unintended consequence of performing worse than it would have using random sampling. This can happen when the model has not learned the target concept well enough to accurately assess the confidence of its predictions. A bad model can steer the active sampling procedure away from the kinds of observations which are likely to be wrong (i.e., the observations we assume will improve the model the most), because the model is overconfident. From an information retrieval point of view, this means that the model should have high recall with regard to selecting which instances it will likely get wrong in order to address the areas of the feature space which result in the most confusion for the model. For more information on theory and application of active learning, we recommend reading [21].

Semi-supervised learning approaches can also leverage model confidence, but they do not explicitly lend themselves to interactivity. The *self-training* algorithm uses an initial model learned using all available labelled data to make predictions on the unlabelled data. The predicted labels of instances with the highest prediction confidence, or likelihood of having a correct prediction, are assumed to be equivalent to ground truth. In theory, incorporating these observations and their predicted labels into the training set and retraining the model will yield higher model performance. With this approach, it is not the recall of the selected samples that matters, but rather the precision. Incorrect predictions will "pollute" the training set if selected. For more information on the theory and application of semi-supervised learning methods we suggest [29]. One potential caveat of using this approach in activity recognition is that each self-training iteration is likely to result in new model behaviours where the patterns of correct and incorrect predictions that the user has come to expect may no longer be valid (Fig. 16.6).

User-Centred Validation Algorithm for Experiment 2

We can modify the validation algorithm in experiment 1 to better understand how we might leverage model confidence for the benefit of an activity recognition system. In algorithm 2, we specifically seek to analyse accuracy across users, among a model's 30 most confident (MC) predictions and 30 least confident (LC) predictions in order to help us make system and user interaction design decisions with regard to active learning, semi-supervised learning, or seamful design approaches.

The purpose of algorithm 2 is to help us understand the expected quality of model confidence for each user using either a personal, hybrid, or impersonal model. Algorithm 2 differs from algorithm 1 in two ways. First, with each user, all data that is not part of the personal training set, $D_{personal}$, is added to a pool of unlabelled personal data, $U_{personal}$. We then record predictions and their respective model confidence on all observations in the $U_{personal}$ dataset with the impersonal, personal and hybrid models. The second difference is that the predictions are now ordered from most confident to least confident predictions. Model confidence in the SciKit-Learn python module is calculated as the mean predicted class probabilities of all decision trees in the forest. For each decision tree, the class probability is the fraction of samples of the same class in a leaf [20].[3] We can think of model confidence as a rough approximation of the probability that our prediction is true or $p(\hat{y} = y)$ where $\hat{y}$ is our activity prediction and y is the actual activity label. With the confidence ranked predictions for a user's unlabelled data, $\hat{Y}_{\theta,user} = \text{argsort}_y\ p(\hat{y} = y)$, we can now assess accuracy with respect to seamful designs which emphasise predictions that are most likely to be correct, $\hat{Y}_{MC} = (y_0, \ldots, y_i)$ (pessimistic presentation) or seamful designs which emphasise predictions that are most likely to be incorrect, $\hat{Y}_{LC} = (y_{n-j}, \ldots, y_n)$ (cautious or opportunistic presentation). Here, i is the cutoff in the confidence ranked predictions where predictions indexed greater than i are no longer trusted to be correct and j is the index where all values indexed greater than j represent the least confident samples that we are interested in evaluating.[4]

Experiment 2: Results and Discussion

In Fig. 16.7, we see that the inner quartile range and mode for accuracy of all models across users shifts upward when we select only the 30 most confident examples and downward for the 30 least confident predictions for each user. As seen in the top panel of Fig. 16.7 while the modal accuracy for the impersonal model shifts up to 83% from 45%, there are still many cases of poor accuracy among these most confident predictions. When considering the results in the context of a pessimistic

[3] The way in which model confidence is assessed can vary in many ways. It can vary depending on how we determine class probability from the model. For example, a K-nearest neighbors algorithm might assess confidence as the average distance to the neighbors of a new observation while an SVM approach might assess confidence as the distance from a new observation to the hyperplane used to separate classes. Model confidence can vary depending on utility functions as described in the second chapter of [21]. It can also vary depending on whether or not evidence is taken into account [22].

[4] In our experiment we take $i = j = 30$ for ease of comparison and exposition, but in practice these cutoff points can vary and be optimised for recall or precision as mentioned earlier in this section.

Algorithm 2 User-Centred Validation Pipeline for Model Confidence in Activity Recognition

```
Let $D$ be all the labelled activity recognition data
    we have available to study
$D_{personal}$ will represent the subset of data representing a user, $u$

for all $u$ in the set of users in $D$
    Let $D_{personal}$ be all data in $D$ where $D_{personal} == u$
    Let $D_{impersonal}$ be all data in $D$ where $D_{personal} != u$


    Let $T_{personal}$ be the subset of $D_{personal}$
        which was sampled through some sampling function,
        $s(D_{personal})$

    Let $U_{personal}$ be the subset of $D_{personal}$
        where all $x$ in $U_{personal}$ is not in $T_{personal}$ to serve as
        the unlabelled dataset.

    Let $T_{hybrid}$ be the training set which combines data from both
        $T_{personal}$ and $T_{impersonal}$ by joining the sets
        through some function $j(personal, impersonal)$

    Let $\theta_{personal}$ be the model trained on $T_{personal}$
    Let $\theta_{hybrid}$ be the model trained on $T_{hybrid}$
    Let $\theta_{impersonal}$ be the model trained on $T_{impersonal}$

    Using $\theta_{personal}$ record predictions ranked by model confidence,
        $\hat{Y}_{\theta_{personal}, user}$ for each instance in $U_{personal}$.
    Using $\theta_{hybrid}$ record predictions ranked by model confidence,
        $\hat{Y}_{\theta_{hybrid}, user}$ for each instance in $U_{personal}$.
    Using $\theta_{impersonal}$ record predictions ranked by model confidence,
        $\hat{Y}_{\theta_{impersonal}, user}$ for each instance in $U_{personal}$.
```

seamful design where we use impersonal model confidence as a way of signaling to the user when they should trust the system, some users will likely benefit, but many others will still find incorrect predictions among even the most confident examples. A self-training approach can also yield poor results considering that many (sometimes all) of the most confident predictions are incorrect for a user. Selecting only the least confident examples, as a system might do in an active learning approach, or to highlight moments of system confusion in a cautious seamful design approach, appears to yield mostly instances which are likely to be wrong, but it is difficult to resolve this with the overall likelihood of the model to make mistakes.

Table 16.1 focuses on likely experiences for individual users and shows those who get the least accuracy from the 30 least confident predictions each model makes. The users represented in rows 1, 4 and 5 of the impersonal model do not get any reasonable accuracy overall or within the most confident examples meaning that the low accuracy

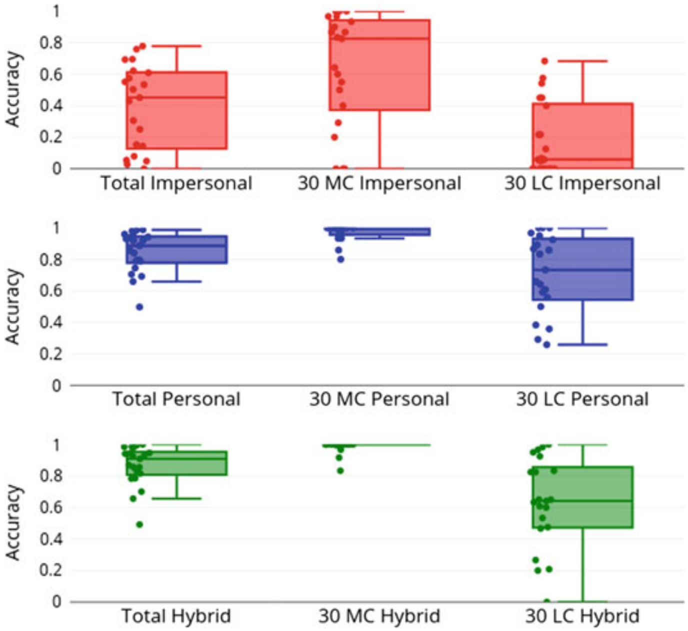

Fig. 16.7 This figure shows model accuracy for impersonal (top), personal (middle) and hybrid (bottom) approaches when we select only 30 most confident (MC) or 30 least confident (LC) and compare them to the model's overall accuracy. Similar to previous graphs, each data point represents a single user. The hybrid and personal models were each trained using 5 personal samples using the same sampling function described in experiment 1

Table 16.1 Impersonal, personal and hybrid model accuracies for all predictions, 30 most confident (MC) predictions and 30 least confident (LC) predictions, among the 5 who received the worst accuracy from the least confident predictions of that particular model

Impersonal			Personal			Hybrid		
All	30 MC	30 LC	All	30 MC	30 LC	All	30 MC	30 LC
0.00	0.00	0.00	0.66	1.00	0.26	0.94	1.00	0.00
0.45	0.82	0.00	0.50	0.80	0.29	0.66	1.00	0.20
0.55	1.00	0.00	0.69	1.00	0.36	0.79	0.99	0.21
0.027	0.50	0.00	0.74	1.00	0.38	0.49	0.83	0.27
0.14	0.20	0.00	0.71	0.86	0.50	0.93	1.00	0.47

is likely due to the model's inability to generalize to this user. However, those users represented in rows 2 and 3 can at least benefit from self-training or by integrating model confidence as pessimistic or cautious seamful design. With a user-centred perspective, we can further explore the cases that stand out to understand whether there is potential for a system design that can at least provide model confidence before requiring manual input from the user.

For both personal and hybrid models, the most confident predictions are highly accurate across users, with no single user getting less than 80% accuracy from these predictions. This means that after a user has done about 50 s of activity labelling, an activity tracking system should consider leveraging self-training to improve accuracy or pessimistic seamful design techniques to help the user understand when the inferred information is most reliable. The least certain examples are generally more accurate than those of the impersonal model, but this is likely due to overall accuracy of the model meaning that even the least confident predictions are still likely to be correct. However, this isn't always the case. A closer look at the user who received the lowest accuracy from the least confident predictions of the hybrid model (also represented in the first row of the hybrid model column in Table 16.1) reveals that this user also received 100% accuracy from their most confident predictions and that overall the accuracy was 94%. This means that the 30 least confident predictions represent all of the incorrect predictions and would have been very helpful for an active learning or cautious seamful design approach.

It's important to note that we chose 30 to be the number of most and least confident examples somewhat arbitrarily to simplify our analysis. In practice, these models will need to be adapted to a stream-based approach in which unlabelled personal instances are observed one-by-one and a decision about whether to query the user for a label will need to be made before the user becomes unaware of exactly what activity they were doing during the 10-second window of time that the unlabelled instance represents. The pool based approach we demonstrate here is representative of some of the earlier approaches to understanding whether active learning is theoretically possible in an activity tracking system [3, 14, 17, 24].

16.4 Discussion

User-centred evaluation has long been a central component of user interface and user experience design. Intelligent systems which aim to provide users with new information that can be used to make better decisions rely on complex and sophisticated machine learning algorithms. These algorithms need to observe human behaviour in order to model it, making them inherently dependent on the human experience even though the algorithms themselves do not directly face the user. To evaluate an algorithm in machine learning, we often have to reduce the expected context of the system (many users with many different and unique styles of activity behaviour) to a problem which is easier to evaluate and optimise. We believe user-centred evaluation can be integrated into the algorithm design process by adapting principles from the International Usability Standard as stated in Sect. 16.2.3 and illustrated in experiments 1 and 2.

During the stage of system development concerning the learning algorithm and a model representation of activity, we can incorporate an understanding of the users, their goals and tasks and their environments to show that not only can the concept of activity be learned using machine learning algorithms, but that the technology can help users in achieving their broader goals that stem from activity tracking. To help guide our understanding of the design challenges facing the development of activity tracking technology, we studied the research regarding current commercial activity trackers from the user perspective. Researchers observe that people have varying motivations for using the technology including maintaining a healthy exercise routine or finding opportunities to build better health habits [10, 23, 27]. Similarly, user lifestyles range from amateur athletes looking to challenge themselves to recent mothers who may be pushing a stroller while exercising. When the users begin to witness the first incorrect predictions, their task shifts from leveraging tracked information to testing the boundaries of the system's capabilities by creating ad-hoc folk tests [27]. These prior studies provide detailed insight into the way users interact with the technology, their purpose for interacting with the technology and the context and environments in which they employ it to best understand the appropriateness of an approach.

This enriched understanding of variability in users guided our development of an extension to a standard leave-one-user-out algorithm that allows us to better understand the variability in user experience from the model perspective. In experiment 1, we saw that users of a system with an impersonal model fall into one of two clusters: one which experiences less than 25% accuracy and one which experiences between 45 and 79% accuracy. With uniform probability across five potential activity classes, the former group will experience performance that is, at best, slightly better than a random guess. These users have the highest benefit-to-burden ratio - meaning that they have the most to gain from their labelling interactions with the system. For the latter group, the model performs better than chance meaning that while it may have learned something about activity as it relates to these users, they will frequently experience incorrect predictions making it difficult for them to utilise the information to make decisions regarding their health. These users will still have a positive benefit-to-burden ratio, but will experience lower gains for each label contributed than users in the first group. In experiment 2, when we look at the group of users with the lowest accuracy in the set of the least confident predictions, we can see two users with 45 and 55% overall impersonal model accuracy who get 82 and 100% accuracy among the 30 most confident predictions made by the model. What this means is that while the impersonal model may present a low benefit-to-burden ratio for a labelling interaction for some users, it can leverage model confidence to lower burden or increase benefit. For example, this model can lower burden by using the self-training algorithm which is likely to increase model accuracy making manual interaction less necessary or increase benefit by leveraging seamful design approaches like the pessimistic presentation of model confidence which fosters greater understanding of the model's behaviour in the user. By simply validating with respect to individual users we can derive much greater insight at the level of algorithm design than we can when validation is agnostic to the individual users who are represented in the test data.

A better understanding of users also helps to address the whole user experience (the third usability standards principle) while making decisions about the underlying algorithmic components of the system. Impersonal models have not only shown that they provide suboptimal accuracy, they also lack the interactivity that users need to calibrate the model to their personal needs, test their understanding of the model's behaviours, and foster the engagement and awareness of tracked information that help users to make better health decisions. These interactive capabilities can be thought of in the context of a calibration phase or in the context of ongoing manual intervention. An initial calibration phase is a necessity of a personal model and would set the expectation that the system will likely not understand scenarios where the user has not provided examples. A hybrid model may also require an initial calibration phase though there is the possibility that for some people the model is at least modestly capable of making predictions without it. Continued model personalisation that leverages model confidence could "unlock" after the system has reason to believe that the model will perform at least adequately for the user. For example, the data labelled during the calibration phase could first be used as a personal test set. If the model fails to meet a certain threshold of accuracy for the calibration data, then the data can be incorporated into the training set so that the model can be retrained and a notification for new calibration data can be made randomly at some point in the near future. If the model exceeds a threshold for accuracy in predicting the calibration data, then the model confidence features can be unlocked since it can at least be confident in its ability to select observations that it is likely to predict correctly or incorrectly.

16.5 Future Directions

While we know that users are interested in interactive design components, designing interactions and interfaces for intelligent systems is complicated by the behaviour of the underlying model given the way it handles input from varying users. Furthermore, dynamic models, such as those used in model personalisation which continue to learn after being deployed to the user, will learn new discriminative patterns over time. How users will feel about this learning process remains an open question that will be difficult yet important to study empirically. Patterns of model behaviour that the user noticed early in their use of the system may no longer hold after the model further adapts to the user's personal behaviour. Studies of intelligent user interfaces have shown how interactive features allowing the user to do things like roll back model changes can give the user the level of control they need to maintain a reasonably working system [11].

From a modelling perspective, we know that physical activity tends to be fairly consistent over short periods of time meaning that the likelihood of walking at time, t, is heavily dependent on whether we were walking at time $t - 1$. Much of the research that we cite and conduct in this paper does not take this into consideration and this is mostly because of the added level of complication it adds to analysis. That said, it is an important aspect of activity recognition that should be studied in context with model personalisation and interactive approaches to activity tracking system design.

[1, 2, 18] are examples of research that we know of in activity recognition that are studying active learning in a stream-based temporal context.[5]

To address the shortcomings of impersonal models, some researchers are studying the “population diversity” problem in which impersonal datasets can include a wide variety of activity behaviour, much of which is irrelevant to the activity behaviour of many individual users. With a better understanding of the *population diversity* of impersonal datasets and how the data can better complement personal data, we may be able to better utilise the impersonal data when combining it with personal data for a hybrid model. For example, [13] have devised methods for comparing an individual end-user’s data to the individuals in the impersonal dataset in an effort to filter other users who have data which is likely to be beneficial to the end-user’s hybrid model. [9] aim to address population diversity by combining the instances of a particular class from the impersonal and personal datasets which were found to be similar through a clustering approach. This can also be thought of as a transfer learning task in which some of the knowledge learned from training a model on impersonal data (e.g., clusterings of which users have similar activity behaviors can be *transferred* to a *personal* model). Future work should consider using neural network algorithms for transfer learning which learn lower level features that are more likely to generalise accurately to new users in a transfer learning task [28]. For a comprehensive survey of transfer learning research in the activity recognition space refer to [8].

Additionally, we can expect a kind of “concept drift” where activity behaviour changes either suddenly due to injury or slowly due to aging causing the discriminatory patterns learned by the model at one point in time to lose its predictive accuracy. Whether adaptive modelling approaches alleviate or exacerbate this effect is an open question. Future work should seek to apply user-centred evaluation to understand how models of activity recognition which are adaptive, temporal and stream-based could be used in interactive and seamfully designed activity tracker systems and how they will behave over extended periods of time.

16.6 Conclusion

Designing intelligent systems often begins with the most novel component, the learning algorithm, independent from other components like the user interface. As a result, the objective when optimising the algorithm (e.g., to minimise errors as much as possible) is often misaligned with the user’s goal (e.g., to understand patterns in active behaviour in the case of activity trackers). We demonstrate how user experience research can help inform model optimisation and selection so that evaluation processes which are more user-centred can be developed and integrated into the development process. These user-centred evaluation methods can highlight problematic patterns which help with selecting a model which addresses the whole user experi-

[5] The temporal component also introduces the added complexities addressed by the online learning and incremental learning research within machine learning.

ence. User-centred evaluation can also highlight opportunities for seamful design. Using this process we found impersonal models for activity recognition to be problematic because they present poor model accuracy (<25%) for many and mediocre model accuracy (45–79%) for the rest. Additionally, we define a benefit-to-burden ratio metric as the ratio of the amount of expected benefit to the user and their system (mostly, but not exclusively with respect to model performance) to the amount of expected burden to the user (mostly, but not exclusively with respect to the amount of interaction). Using this, we find that most models for activity recognition (based on random forest regression trees) which perform with better than 45% accuracy are capable of leveraging model confidence, appear capable of selecting predictions which are likely to be incorrect and predictions which are likely to be correct. This representation of model confidence can be leveraged for model personalisation approaches such as self-training and active learning as well as seamful design features such as those that present predictions pessimistically (only those which are likely to be correct) or cautiously (only those which are likely to be incorrect).

References

1. Abdallah, Z.S., Gaber, M.M., Srinivasan, B., Krishnaswamy, S.: StreamAR: incremental and active learning with evolving sensory data for activity recognition. In: 2012 IEEE 24th International Conference on Tools with Artificial Intelligence, vol. 1, pp. 1163–1170 (2012)
2. Abdallah, Z.S., Gaber, M.M., Srinivasan, B., Krishnaswamy, S.: Adaptive mobile activity recognition system with evolving data streams. Neurocomputing **150**, 304–317 (2015)
3. Alemdar, H., van Kasteren, T., Ersoy, C.: Using active learning to allow activity recognition on a large scale. In: Ambient Intelligence, pp. 105–114 (2011)
4. Bao, L., Intille, S.: Activity recognition from user-annotated acceleration data. In: Pervasive Computing, pp. 1–17 (2004)
5. Chalmers, M.: Seamful design: showing the seams in wearable computing. In: Proceedings of IEE Eurowearable'03, vol. 2003, pp. 11–16. IEE (2003)
6. Chalmers, M., MacColl, I.: Seamful and seamless design in ubiquitous computing. In: Workshop at the Crossroads: The Interaction of HCI and Systems Issues in UbiComp, vol. 8 (2003)
7. Choe, E.K., Abdullah, S., Rabbi, M., Thomaz, E., Epstein, D.A., Cordeiro, F., Kay, M., Abowd, G.D., Choudhury, T., Fogarty, J., Lee, B., Matthews, M., Kientz, J.A.: Semi-automated tracking: a balanced approach for self-monitoring applications. IEEE Pervasive Comput. **16**(1), 74–84 (2017)
8. Cook, D., Feuz, K.D., Krishnan, N.C.: Transfer learning for activity recognition: a survey. Knowl. Inf. Syst. **36**(3), 537–556 (2013)
9. Garcia-Ceja, E., Brena, R.: Building personalized activity recognition models with scarce labeled data based on class similarities. Ubiquitous Computing and Ambient Intelligence. Sensing, Processing, and Using Environmental Information. Lecture Notes in Computer Science, pp. 265–276. Springer, Cham (2015)
10. Harrison, D., Marshall, P., Bianchi-Berthouze, N., Bird, J.: Activity tracking: barriers, workarounds and customisation. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '15, pp. 617–621. New York, NY, USA (2015)
11. Kulesza, T., Burnett, M., Wong, W.K., Stumpf, S.: Principles of explanatory debugging to personalize interactive machine learning. In: Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI '15, pp. 126–137. ACM Press, New York (2015)

12. Kwapisz, J.R., Weiss, G.M., Moore, S.A.: Activity recognition using cell phone accelerometers. ACM SigKDD Explor. Newsl. **12**(2), 74–82 (2011)
13. Lane, N.D., Xu, Y., Lu, H., Hu, S., Choudhury, T., Campbell, A.T., Zhao, F.: Enabling large-scale human activity inference on smartphones using community similarity networks (csn). In: Proceedings of the 13th International Conference on Ubiquitous Computing, pp. 355–364. ACM, New York (2011)
14. Liu, R., Chen, T., Huang, L.: Research on human activity recognition based on active learning. In: 2010 International Conference on Machine Learning and Cybernetics, vol. 1, pp. 285–290 (2010)
15. Lockhart, J.W., Weiss, G.M.: The benefits of personalized smartphone-based activity recognition models. In: Proceedings of the 2014 SIAM International Conference on Data Mining, pp. 614–622. SIAM (2014)
16. Lockhart, J.W., Weiss, G.M., Xue, J.C., Gallagher, S.T., Grosner, A.B., Pulickal, T.T.: Design considerations for the WISDM smart phone-based sensor mining architecture. In: Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data, pp. 25–33 (2011)
17. Longstaff, B., Reddy, S., Estrin, D.: Improving activity classification for health applications on mobile devices using active and semi-supervised learning. In: 2010 4th International Conference on Pervasive Computing Technologies for Healthcare, pp. 1–7 (2010)
18. Miu, T., Missier, P., Pltz, T.: Bootstrapping personalised human activity recognition models using online active learning. In: 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), pp. 1138–1147. IEEE (2015)
19. Patel, M.S., Asch, D.A., Volpp, K.G.: Wearable devices as facilitators, not drivers, of health behavior change. JAMA **313**(5), 459–460 (2015)
20. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., others: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
21. Settles, B.: Active learning literature survey. University of Wisconsin, Madison, vol. 52(55–66), p. 11 (2010)
22. Sharma, M., Bilgic, M.: Evidence-based uncertainty sampling for active learning. Data Min. Knowl. Discov. **31**(1), 164–202 (2017)
23. Shih, P.C., Han, K., Poole, E.S., Rosson, M.B., Carroll, J.M.: Use and adoption challenges of wearable activity trackers. In: iConference 2015 Proceedings (2015)
24. Stikic, M., Van Laerhoven, K., Schiele, B.: Exploring semi-supervised and active learning for activity recognition. In: 12th IEEE International Symposium on Wearable Computers (ISWC2008), pp. 81–88 (2008)
25. Weiser, M.: Some computer science issues in ubiquitous computing. Commun. ACM **36**(7), 75–84 (1993)
26. Weiss, G.M., Lockhart, J.W.: The impact of personalization on smartphone-based activity recognition. In: AAAI Workshop on Activity Context Representation: Techniques and Languages, pp. 98–104 (2012)
27. Yang, R., Shin, E., Newman, M.W., Ackerman, M.S.: When fitness trackers don't 'fit': end-user difficulties in the assessment of personal tracking device accuracy. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '15, pp. 623–634. New York, NY, USA (2015)
28. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems, pp. 3320–3328 (2014)
29. Zhu, X., Goldberg, A.B.: Introduction to semi-supervised learning. Synth. Lect. Artif. Intell. Mach. Learn. **3**(1), 1–130 (2009)