HHS Public Access

Author manuscript

*IEEE J Biomed Health Inform*. Author manuscript; available in PMC 2020 July 01.

Published in final edited form as:

*IEEE J Biomed Health Inform*. 2019 July ; 23(4): 1585–1594. doi:10.1109/JBHI.2018.2869779.

# Classifier Personalization for Activity Recognition using Wrist Accelerometers

**Andrea Mannini** and

BioRobotics Institute, Scuola Superiore Sant'Anna, Pisa, Italy.

**Stephen S. Intille**

College of Computer and Information Science and Bouvé College of Health Sciences, Northeastern University, Boston, MA.

## Abstract

Inter-subject variability in accelerometer-based activity recognition may significantly affect classification accuracy, limiting a reliable extension of methods to new users. In this work we propose an approach for personalizing classification rules to a single person. We demonstrate that the method improves activity detection from wrist-worn accelerometer data on a four-class recognition problem of interest to the exercise science community, where classes are *ambulation*, *cycling*, *sedentary*, and *other*. We extend a previously published activity classification method based on support vector machines so that it estimates classification uncertainty. Uncertainty is used to drive data label requests from the user, and the resulting label information is used to update the classifier. Two different datasets – one from 33 adults with 26 activity types, and another from 20 youth with 23 activity types – were used to evaluate the method using leave-one-subject-out and leave-one-group-out cross validation. The new method improved overall recognition accuracy up to 11% on average, with some large person-specific improvements (ranging from −2% to +36%). The proposed method is suitable for online implementation supporting real-time recognition systems.

### Index Term

Active Learning; Activity Recognition; Incremental Learning; Personalization; Support Vector Machines; Wearable Sensors

## I. INTRODUCTION

ACTIVITY recognition using wearable sensors can support scientific measurement of physical activity and sedentary behavior for research in health and medicine [1]. Many scientific studies using "objective," accelerometer-based sensors in lieu of self-report have participants in experiments wear a single sensor on the hip, above their clothing. In this location, a sensor measures overall body motion, such as ambulation. Unfortunately, the sensors have to be removed at night and when changing clothes, and some participants find

andrea.mannini@santannapisa.it.

attaching them to the hip to be inconvenient or aesthetically undesirable [2]. As a result, missing data are common. With the introduction of sensors capable of saving accelerometer data at 60+ Hz, and the promise of development of activity recognition algorithms aimed at interpreting those data to differentiate gesturing from true ambulation (e.g., [3, 4]), researchers began collecting data with a single sensor placed at the wrist. Even large, influential studies such as the UK Biobank and the U.S. National Health and Nutrition Examination Survey (NHANES) changed data collection from the hip to the wrist locations [5, 6]. The introduction of smartwatch devices, nearly all of which include accelerometers, has further intensified interest in recognizing activity from wrist data.

Differentiation of behavior from wrist data is challenging due to gesturing and differences in how individuals perform activities. In our prior work, we observed that wrist-based classifiers trained using leave-one-subject-out (LOSO) cross validation performed well overall, but generalized inconsistently across different people [7, 8]. The problem is that in LOSO training, models are (appropriately) trained without any data from the left-out individuals; but this restricts generalization to people who may perform activities in an uncommon way. The aim of this work is to use a small amount of *personal* data to infuse a general model with enough subject-specific information to improve performance, but without leading to overfitting or unrealistic expectations about the amount of subject-specific training data that must be obtained.

## II. PRIOR WORK

The need for machine learning classification personalization has been widely discussed in domains such as automatic handwriting recognition [9–14], speech recognition [15], and hand gesture recognition [16]. In all these cases, personalizing the classifier – that is, adapting a generalized classifier using some person-specific training data – improves classification accuracies while limiting the need to acquire a large, person-specific dataset from every individual. In the field of human activity recognition from accelerometer data, the automatic personalization of classifiers is still under investigation [17].

The impact of person-specific data was observed in one study that considered a 59-participant dataset, where each participant performed six classes of activities (*sit*, *stand*, *lie*, *cycle*, *walk* and *run*) [18]. Acceleration data were acquired using a smartphone, and three validation conditions were compared: individual (cross-validation training using data from a single subject), leave-one-subject-out (LOSO) (cross-validation training using data from participants different from the participant being tested), and hybrid (cross-validation training using all available data). The best performance was obtained using cross-validation and data from single subjects [19]. This is not surprising, because the training examples are likely to closely mirror the test examples; the downside is that a large amount of training data must be gathered from every individual to make robust person-specific classifiers that do not over fit the data [9], and gathering training data is a burdensome task. An alternative approach is to limit the variability of expected data, clustering available data for specific groups of subjects according to their age, gender, weight or health conditions, and training group-specific classifiers [20]; many different classifiers are created. The assumption is that similar people

perform activities in similar ways, but that assumption may not hold true, especially when activities must be inferred from highly-variable, wrist-worn sensor data.

Alternatively, with *personalization*, a classifier trained on a general population is adapted to the particular person it is designed for [21], providing improved classification. By starting the customization from a general classifier, far fewer datasets can be used to train a person-specific classifier than if the classifier is trained anew, which is important given the burden and complexity of obtaining person-specific training data. One proposed personalization method uses a decision tree classifier in which classification thresholds are tuned to the specific person using 3–10 min of new user-annotated data [21]. The algorithm was evaluated on a dataset of accelerometer data (wrist and ankle) collected as seven volunteers performed six activities (*sit*, *stand*, *lie*, *cycle*, *walk* and *run*). The personalization improved accuracy by 7.4% with respect to the LOSO result.

Generally, two personalization strategies have been tested in prior work: using semi-supervised learning (SSL) or using active learning (AL). SSL algorithms use unlabeled data to refine classification rules, by associating unlabeled clusters of data with known classification labels [22]; classification accuracy is improved by increasing the size of the training dataset with unlabeled data [23]. AL algorithms, alternatively, request additional labeled training examples when certainty about classification decisions is low, attempting to maximize acquisition of informative samples while minimizing annotator intervention/burden [24]. In AL, but not SSL, user intervention is required to extract labels.

AL and SSL are widely used for improving human activity recognition from accelerometer data [23–26], but typically without adapting the classification rules to the specific user. For instance, AL can reduce the amount of labeled data needed for training a sparsely-annotated dataset; an algorithm can automatically choose a meaningful subset of data to label and request those labels, instead of asking for the full dataset to be labeled [24, 27]. AL can also improve algorithm reliability with respect to changes in time evolution of data streams, i.e. giving the algorithm the capability to detect changes in the observed data over time, and then enabling an adaption of its classification rules to accommodate those changes [25, 26].

In more recent work, AL vs. SSL strategies have been compared when using a smartphone as a data logging platform to collect accelerometer data for a three-class activity discrimination task (*standing still*, *walking* and *running*, 30 min each by 32 participants) [28]. As expected, the information acquired from user intervention (i.e., labeling) in AL allowed the AL strategies to significantly outperform SSL strategies on this task. However, the studies by Longstaff et al. can be extended by considering that a limitation of the C4.5 decision tree algorithm used, is that the entire training dataset must be stored in the memory of the system for retraining to achieve personalization. In addition, the study was based on a dataset including only the three visually distinct activities of walking, running and standing still. Moreover, the effect of using different values for the classification uncertainty threshold, which impacts the volume of user requests needed to execute AL (and therefore user burden), was not evaluated.

In this work we explore whether a personalization method improves an existing solution for human activity recognition by testing how personalization impacts performance of a four-class activity recognition solution presented in previous work [8]. Because others have shown AL outperforming SSL [28], we employed the same AL-based personalization approach used in prior work and left testing SSL algorithms for future work. Our goal was to develop a practical implementation of a personalization strategy that can be implemented online, with low computational and memory requirements.

Unlike previous studies, we evaluated the approach using two datasets acquired from two different age groups of users: 33 adults (*adult dataset*) and 20 youth (*youth dataset*); these participants performed 26 and 23 different types of activities, respectively, which were grouped into four classes for a total of ~4800 min of data (~ 90 min/person). The personalization algorithm was evaluated using two different cross-validation approaches: leave-one-subject-out (LOSO) and leave-one-group-out (LOGO). In LOSO, the generalized classifier was obtained by training on all participants' data except the one participant being tested. In LOGO, the generalized classifier was obtained by training the classifier using data from the alternative group (i.e., a classifier was trained on the adult dataset and tested on youth participants or trained on the youth dataset and tested on adult participants). Differences between training and testing data are most significant in the LOGO condition due to the differences in the age of the participants and the activities they performed.

Our classifier personalization approach is suitable for online, incremental learning (IL) [29]. In IL, the learning mechanism is adjusted every time new information is available; this differs from batch learning, in which the training occurs only once, using all available data. To our knowledge, the application of IL in the personalization of accelerometer-based activity recognition methods has not been extensively tested yet; an exception is a recent study by Siirtola et al. in which IL-capable algorithms for personalizing the recognition of seven activities (walking, sitting, standing, jogging, biking, up/down stair walking) were tested using a dataset acquired on 10 adults [30]. Other studies use IL to incrementally refine activity classification rules jointly with AL-based strategies, but its application is not on personalizing classification rules; it is aimed at the selection of a subset of the dataset to be labeled to speed up the training phase instead [25, 26].

## III. MATERIALS AND METHODS

### A. Study organization

Our prior work demonstrated activity recognition using SVM classifiers and wrist-worn accelerometer data from adult [7] and youth [8] data. Four activity classes were recognized (*ambulation*, *cycling*, *sedentary*, and *other*) based on windowed accelerometer of data, using nine frequency- and time-based features. Results showed variability of classification accuracy across different subjects, with recognition rates on single subjects varying from 80.8% to 96.5% in youth and from 70.7% to 95.0% in adults. To limit that variability, in this work a personalized activity recognition algorithm was developed and evaluated.

Both of our previous studies followed the typical LOSO and LOGO approaches (the steps outside the shaded box in Fig. 1) [7, 8]. In LOSO, a generalized classifier was trained using

all available labeled data except the data from one participant. The classifier broke the data into 12.8s labeled windows of time, computed features using the raw accelerometer data, and then used the batch of labeled feature vectors to train a non-personalized classifier. That fixed classifier was evaluated against the data from the one left-out participant. This process was repeated for all participants, averaging the results. In LOGO, a generalized classifier was trained using the batch of available data from one group (adults or youth), fixed, and then tested on all available people in the other group, averaging the results.

Unlike that prior work, the classification validation step here proceeded in an online, incremental learning fashion. The algorithms were initially trained as before, using all available data (depending upon the LOSO or LOGO case). Then, windows of test set data were presented to two different versions of the algorithm — the fixed non-personalized version, and the adapting personalized version — one window at a time. Windows were classified using both versions, and then results were compared. The new personalization algorithm uses distance to the SVM decision boundaries as a proxy for classification certainty, to be described in more detail in Section III.C. Classification results deemed "uncertain" (i.e., close to a decision boundary, $> Th_i$) were selected and then included in a new training set along with support vectors retained from the general classifier, thereby leading to a refinement of classification rules of the personalized classifier before the next window was evaluated (details in Section III.D). Data that entered in the training was never used for testing: both cross-validation approaches exclude data from the subject (or group) being tested. Then, as reported in Fig. 1, the personalization occurs on data from the subject being tested only after the classification step.

All the processing steps described were implemented in Mathworks Matlab (version 2017a), and the software and datasets used here are available online (http://mhealthgroup.org/datasets).

## B. Datasets and features

The datasets used in this paper are described in prior work [7, 8]. Adults and youth, respectively, performed a set of daily activities in a lab environment while wearing a custom triaxial accelerometer [31] attached to the wrist. The wrist sensor was placed on the dorsal side of the dominant wrist midway between the radial and the ulnar processes. The custom accelerometers [31] were used because they are small, thin, and lightweight devices (43 × 30 × 7 mm, 13 g). These features make them particularly suitable for long-term physical activity monitoring studies, where mobile phones are used for data collection. Raw acceleration data (range ±4 g) were acquired at 90 Hz and sent using the Bluetooth wireless protocol to a smartphone

Participants performed a guided sequence of laboratory-based physical activities and common daily activities. Activities were annotated during the execution of tasks using a voice recorder, and then timestamps on the voice recording were used to annotate start/stop times for specific activities being observed. Data and annotation were synchronized using custom software [31]. The list of available activities in the youth dataset compared with activities annotated in the adult dataset is summarized in Table I. To remain consistent with our previous work, activities were grouped into four broad classes: *sedentary*, *cycling*,

*ambulation* and *other* activities The *other* class included activities that were not sedentary and that were done in the upright position [7, 8].- These classes were originally selected because they broadly cluster the main activities of daily living, and knowledge of which activity category is active could improve energy expenditure estimation, a topic of great interest to researchers in exercise science [7, 8, 32]. The data collection procedure captured natural variability in how activities are performed within each broad class. Some errors in annotation at activity transitions may have occurred due to reaction time when labeling; these times were accounted for by discarding one window of data across transitions. We also followed prior work and discarded unreliable walking labels identified using ankle data acceleration, which was also available to us [7, 8].

Raw acceleration signals (x, y, z) at each time point were combined into a single value by computing a signal magnitude vector (SM), where $SM = \sqrt[2]{acc_x^2 + acc_y^2 + acc_z^2}$ The SM was used to limit impact of sensor orientation. Data were broken into 12.8 s non-overlapping windows. This window size was proposed by Zhang et al. [4] and also applied in our previous studies [7, 8]. Although prior work has shown that other window sizes (e.g., 4s) can be used with only modest degradation of performance [7], here use of the same window length value as in the prior studies allows a direct comparison of results. Features were computed using SM data from each window, again following prior work [8]. Temporal and frequency features consisted of mean value, standard deviation, maximum value, range, first dominant frequency, ratio between the power at the dominant frequency and the total power, ratio between the power at frequencies higher than 3.5 Hz and the total power, and two signal fragmentation features (i.e., the number of high activity samples in the window, and the number of activations episodes in the window, see [8]). Signal fragmentation features that capture the level and duration of activity bursts within the window were also computed using the SM, as in previous work [8].

## C. Classification and uncertainty estimation

SVM classifiers were used for the supervised learning classification tasks (using a radial basis function kernel) [33]. We used the SVM implementation from the LibSVM toolbox [34]. SVM classifiers are desirable because the optimization criteria are convex, which implies that a global optimal solution exists. Another advantage of SVM classifiers is that they can provide an estimate of class-conditional probability in addition to the classification outcome. To do this, the LibSVM posterior probability estimate routine was used, [34]. In this method a logistic regressor is cascaded to the SVM's output to estimate posterior probability, [35]. Such soft-assignment permits evaluation of classification uncertainty by estimating the probability of having a particular classification for each data window, where

$$\left(p_{\text{sedentary}} + p_{\text{other}} + p_{\text{cycling}} + p_{\text{walking}}\right) = 1$$

[34]

This probability estimate allows us to mark a classification outcome as uncertain if none of the probabilities exceed a fixed threshold, defined as:

$$Th = \frac{1}{N} + AL_{th} = 0.25 + AL_{th}$$

where N is the number of considered classes (so 1/N indicates the random guess) and $AL_{th}$ is the active learning threshold considered. The behavior of the algorithm varying $AL_{th}$ value from 0 (Th = 0.25, corresponding to a random guess) to 0.7 (Th = 0.95) was evaluated. As $AL_{th}$ is increased, the number of labeled example requests from the algorithm will also increase.

The parameters C = 128 and $\gamma = 0.0625$ of the classifiers were retained from previous work; the parameters were not tuned to the datasets [7, 8].

### D. Classifier adaptation

A classifier is adapted to a specific person when the classification uncertainty for a window of time is deemed too high (i.e., above a fixed threshold); to adapt, the algorithm requests an activity label from the person for that window of time and then updates the model. As previously stated, we prefer to adapt the existing classifier to the new data, instead of training a classifier anew, in order to avoid saving all prior training data in a real-time system. Several well-tested incremental learning algorithms exist for SVM classifiers amenable for use in a personalized activity recognition system [10, 36–41]. Our implementation is based on one of those algorithms that was originally proposed for handwriting recognition [10] and that have limited memory requirements (i.e., does not require storing the full training set). Personalization can be performed by considering newly available windows (actively labeled when their classification was too uncertain) jointly with a subset of the training set that is the support vectors of the previously trained classifier. The classifier can be modified using new examples without keeping all prior examples in memory, as is required with other methods (e.g., the Longstaff et al. method [28]). By setting the active learning threshold at different levels, different levels of uncertainty are tolerated, changing the degree of personalization and the degree of burden.

### E. Validation approaches

Given that two different groups of users were considered in this work (i.e., adults and youth), we tested using both the LOSO and LOGO training approaches. In LOGO, which is a significantly more challenging recognition task than LOSO, a classifier is trained on one group (e.g., adults) and then tested on a different group (e.g., youth). This approach allowed us to test not only how a general classifier adapted to a similar population with the same activity set (LOSO), but also how a general classifier adapted to a different population with a somewhat different activity set (see Table I, adapted from prior work [8]).

### F. Output evaluation

To evaluate uncertainty rejection and personalization, both accuracy and the average F1 score were evaluated. The F1 score is the harmonic mean of precision and recall and it is a measure of the quality of binary (two-class) classifications. Personalized and non-personalized classifiers were compared on a given window using the same test data.

Personalization of classification rules was done *after* the classification step, to make sure that no data ended up in both training and test datasets in either tested condition.

## IV. RESULTS

The results obtained by varying $AL_{th}$ from 0 (no personalization) to 0.7 (uncertainty threshold Th = 0.95) are summarized in Table II and Fig. 2. As $AL_{th}$ increases, so does the percentage of the incoming new data that generate label requests. On the whole, LOSO overall accuracy modestly improved from 88.6% without personalization to 89.5% with personalization, whereas LOGO accuracy improved more substantially from 71.7% to 84.6%, significantly reducing the variability of results across different participants. The achieved person-specific improvement in the LOSO test is modest, only 1% on average when $AL_{th} = 0.7$. In the LOGO test, however, it ranges from 1.3 to 11.1 percentage points on average, depending upon $AL_{th}$. Labeling 22.6% of data leads to improvements for some individuals of up to 24.8 percentage points.

Both accuracies and F1 scores (Fig. 2) increase when personalization is introduced, growing with the fraction of uncertain data windows for which a label is requested to refine classification rules. Such requests provide information but at the cost of introducing burden in an online system.

Table III reports the classification confusion matrices without personalization, with $AL_{th} = 0.4$ (Th of 0.65) for LOSO and $AL_{th} = 0.3$ (Th of 0.55) for the LOGO case. Those values of $AL_{th}$ have been chosen because they may balance performance and number of requests. They result in requests for labels for 9.6% of LOSO data and 22.6% of LOGO data.

Fig. 3 details accuracy results for the four classes. In particular, Fig. 4 (part b) shows a detailed view of the distribution of the accuracy improvement for the four different activity classes at different threshold levels for each of the two cross-validation approaches. Finally, Fig. 4 shows the results of a test using LOGO with $AL_{th} = 0.3$ in terms of accuracy improvement due to personalization and amount of label requests in time, on a window-by-window basis. This figure allows us to evaluate the time evolution of the personalization procedure by showing how fast the accuracy improves (part a) and how many label requests are necessary in time during the procedure (part b). The two upper plots show the time evolution of the accuracy improvement obtained by the LOGO-personalized classifiers with respect to the LOGO-general classifiers. The two lower plots show the ratio of the requested labels with respect to the available labels so far. In addition to median results (solid-lines) data from all participants are reported (dashed lines) and allow the assessment of the variability of personalization effects across subjects.

## V. DISCUSSION

The personalization approach improves classification performance of the LOSO and LOGO conditions, albeit modestly in LOSO. Not surprisingly, as more requests for person-specific labeled data are made and the amount of labeled data provided increases, results improve. The downside to making such requests in a real system is that more effort from users to label their own data is required. In this regard, the obtained improvements may not always justify

the labelling effort requested of the user; minimizing the burden of that effort will be the object of future research. In fact, as reported in Table II and in Fig. 2 (b), the personalization did not affect all participants in the same way. Depending upon the threshold, it positively impacted between 2% and 76% of participants in the LOSO test, and between 32% and 98% of participants in the LOGO test, with peaks of improvements that reached +35.6% for a participant in the LOGO test. The participants who benefited most may have had a slightly different way of moving than the training pool; thus, personalization training might have been more beneficial. Moreover, the ratio of participants showing a degrading effect of personalization is low (1% to 22% in LOSO, 2% to 8% in LOGO), with the worst results degrading classification accuracy by −1.9%. Consider a fixed value of $AL_{th}$ such as 0.4 in the LOSO validation test. Whereas the average change in accuracy due to AL personalization was only +0.6%, one of the participants showed a change of +2.9% and another had a decrease of 0.7%. The same threshold value in the LOGO case resulted in an average improvement of +8.9%, but performance for one person improved +31.3% while decreasing for another by 0.4%. If most people perform activities similarly, personalization will only improve results for the outliers, but for those outliers, improvement could be substantial. Recognition in the LOGO validation test is more challenging, especially for youth data, because the activities differ somewhat in the two datasets (see Table I). For example, the youth dataset includes sport activities; classifiers trained on adult data without these activities perform poorly on the youth dataset, and this is notable in Fig. 3, which shows that *other* and *cycling* classes are those that benefit most from the personalization. Such differences between activities in datasets will be common when applying algorithms in real systems, where individuals will sometimes perform new activities that were not included in training data. Therefore, rather than relying on “black box” algorithms that are trained ahead-of-deployment but then do not adapt, systems will likely need to provide users with a way to label data that cannot be properly processed by the classifier, i.e. recognized with low uncertainty, and adapt classification rules to include the new, labeled examples.

In LOSO, the number of windows selected at each threshold level was similar between adult and youth data. In LOGO, however, a higher percentage of youth label requests, versus adult, were made at the same threshold levels. This change results from the need to gather training data on the new activities not already represented in the training set.

In general, results confirm that the more different the information of the current data window is from the examples used to train the classifier originally, the higher the accuracy gain after personalization. Fig. 4 demonstrates that the classification accuracy is significantly improved in the personalized version of the classifier and that the improvement grows in time, especially for youth data that included additional activities such as sports that were not included in adults data. A limitation of the study made apparent from this figure is that the personalization requires a significant amount of data. This is evident in Fig. 4b because the available amount of data is not sufficient to show an observable reduction of the classification uncertainty in time for all participants (i.e., the number of label requests do not always decrease in time). Ideally, assuming that windows of different activities reach the classifier in random order, as happens in this test, an incremental adaptation of classification rules should require larger efforts at the beginning of the procedure and quickly reduce the number of requests later in time. In the data shown in Fig. 4, however, only a small number

of participants would experience this (see that only some of the dashed lines in Fig. 4b decrease over time). Even though a significant improvement in accuracy can be obtained with respect to the non-personalized classifier, the gain is achieved with a significant participant-labeling effort that adapts relatively slowly, not in one more condensed bust of labeling. Future research could target this limitation and explore mechanism that might allow users to more efficiently provide feedback for personalization. Such work might require the user to better understand what the model is actually doing; the model may need to provide an explanation of a decision. The user may need to use that explanatory information to quickly adapt the model by explaining *why* a provided activity label is what it is. In short, rather than developing algorithms that must learn all relationships from data, some algorithms may be provided with common sense knowledge from users, and only use data to refine and tailor user-specific models.

Working with IL strategies always leads to the plasticity/stability dilemma that establishes the tradeoff between catastrophic interference (aka “forgetting”) and the ability to incrementally and continually accommodate new knowledge in the future whenever new data becomes available [42]. In our particular context, the plasticity should be associated to the fast adaptation capability of the classifier to describe the data of a new user. Reducing plasticity freezes the classifier performance, limiting the number of requests in time. In this work, the online simulation tests did not change the plasticity of the classifier. Such variation could be considered in future work by keeping in memory old support vectors that are now discarded, and then using them during the incremental training process. Further, using variable, instead of fixed, SVM parameters, or choosing subsets of the support vectors to be used in the personalization instead of retaining all of them from the initial training could also play a role in improving plasticity. Such tests are left to future work. However it has to be noticed that the choice of using support vectors instead of the full dataset to run incremental learning allowed us to reduce both computational time for training and memory requirement to store the training data, in fact the number of support vectors in the LOSO case was ~3000 windows instead of the ~12800 elements of the training set whereas in the LOGO case, it was ~1500, with a training set of ~4500 windows for youths and ~8500 windows for adults.

Previous studies in classifier personalization for activity recognition cannot be directly compared to this work, mainly because the datasets used are different (different user groups, number of participants, sensor locations). This work represents the first attempt at running AL-based personalization using data from two different age groups of users performing somewhat different sets of activities, running LOSO and LOGO validation tests, and including ~1.5 hours of labeled activity monitoring data per person. Our results are consistent with results from some prior work showing that when initial accuracy was not high (e.g., <80%), personalization is more effective [28]. In that work an overall improvement of ~12% was obtained when starting from a non-personalized classifier with 76% accuracy, whereas the personalization effect was negligible when starting from a 90% accurate classifier. In the work by Parrka et al., a 7.4% improvement in accuracy after personalization was obtained, cross-validating data (2 youth and 5 adults) using a LOSO approach, without including AL strategies [21].

## VI. CONCLUSION AND FUTURE DIRECTIONS

In conclusion, a previously proposed method for activity recognition using wrist worn accelerometers was extended by validating a methodology to personalize classification rules. Existing classifiers trained on people different from the person being tested can be adapted to the new user, obtaining an improvement in classification performance. Moreover, as it was confirmed in the LOGO tests, more variations on ways of performing activities can be associated with, and then recognized by, the existing class models through this procedure.

The viability of an online implementation of the proposed method is confirmed by simulations. The proposed personalization approach does not require the memory to store the full training set because only support vectors computed from prior training data must be retained during online use. Future work should explore ways to tune the plasticity of the incremental learning solution. Future work might also explore the use of SSL approaches to refine trained classification rules without user intervention and to speed up the personalization procedure. Moreover a temporal filtering strategy could be included to permit the user to label fewer numbers of incoming windows of data, thereby easing user burden, [43].

The proposed method could find application in several fields, not limiting to activity recognition: the suggested methodology can be applied to all SVM-based classification problems that are affected by a lack of generalization capabilities across different groups of users.

## ACKNOWLEDGMENT

The authors thank Drs. William Haskell and Mary Rosenberger, who assisted with collection of data used in this work.

This paper was submitted on May 20th, 2018. This study was made possible, in part, by funding from the National Heart, Lung and Blood Institute, National Institutes of Health award 5UO1HL091737 and by the Italian Ministry of Education, Universities and Research within the FARE framework for excellence in research (ARLEM project, R16H2KJRHA).

## REFERENCES

[1]. Bussmann J, Martens W, Tulen J, Schasfoort F, Van Den Berg-Emons H, and Stam H, “Measuring daily behavior using ambulatory accelerometry: the Activity Monitor,” Behavior Research Methods, Instruments, & Computers, vol. 33, pp. 349–356, 2001.

[2]. Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, and McDowell M, “Physical activity in the United States measured by accelerometer.,” Medicine & Science in Sports & Exercise, vol. 40, pp. 181–8, 2008. [PubMed: 18091006]

[3]. Yang J-Y, Chen Y-P, Lee G-Y, Liou S-N, and Wang J-S, “Activity Recognition Using One Triaxial Accelerometer: A Neuro-fuzzy Classifier with Feature Reduction,” Lecture Notes in Computer Science, vol. 4740, pp. 395–400, 2007.

[4]. Zhang S, Rowlands AV, Murray P, and Hurst TL, “Physical activity classification using the GENEA wrist-worn accelerometer,” Medicine & Science in Sports & Exercise, vol. 44, pp. 742–748, 2012. [PubMed: 21988935]

[5]. Troiano R and Mc Clain J, “Objective measures of physical activity, sleep, and strength in U.S. National Health and Nutrition Examination Survey (NHANES) 2011–2014,” in 8th Internat Conf on Diet and Activity Methods, Roma, Italy, 2012.

[6]. Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, White T, van Hees VT, Trenell MI, and Owen CG, "Large scale population assessment of physical activity using wrist worn accelerometers: The UK Biobank Study," PloS one, vol. 12, 2017.

[7]. Mannini A, Intille SS, Rosenberger M, Sabatini AM, and Haskell W, "Activity recognition using a single accelerometer placed at the wrist or ankle," Medicine & Science in Sports & Exercise vol. 45, pp. 2193–2203, 2013. [PubMed: 23604069]

[8]. Mannini A, Rosenberger M, Haskell W, Sabatini AM, and Intille SS, "Activity recognition in youth using single accelerometer placed at wrist or ankle," Medicine & Science in Sports & Exercise, vol. 49, pp. 801–812, 2017. [PubMed: 27820724]

[9]. Miyao H and Maruyama M, "Writer Adaptation for Online Handwriting Recognition System Using Virtual Examples," in Document Analysis and Recognition, 2009. ICDAR'09. Internat Conf on, 2009, pp. 1156–1160.

[10]. Matic N, Guyon I, Denker J, and Vapnik V, "Writer-Adaptation For On-Line Handwritten Character Recognition," 1993, pp. 187–191.

[11]. Vuori V, Aksela M, Laaksonen J, Oja E, and Kangas J, "Adaptive character recognizer for a hand-held device: Implementation and evaluation setup," 2000, pp. 13–22.

[12]. Kienzle W and Chellapilla K, "Personalized handwriting recognition via biased regularization," in Proc. of the 23rd Internat Conf on Machine Learning, 2006, pp. 457–464.

[13]. Haluptzok P, Revow M, and Abdulkader A, "Personalization of an online handwriting recognition system," in 10th Internat Workshop on Frontiers in Handwriting Recognition, 2006.

[14]. Brakensiek A, Kosmala A, and Rigoll G, "Comparing adaptation techniques for on-line handwriting recognition," in Document Analysis and Recognition, 2001. Proc. 6th Internat Conf on, 2001, pp. 486–490.

[15]. Leggetter C and Woodland P, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, vol. 9, p. 171, 1995.

[16]. Liu J, Zhong L, Wickramasuriya J, and Vasudevan V, "uWave: Accelerometer-based personalized gesture recognition and its applications," Pervasive and Mobile Comput, vol. 5, pp. 657–675, 2009.

[17]. Lara O and Labrador M, "A survey on human activity recognition using wearable sensors," IEEE Communications Surveys and Tutorials, 2012.

[18]. Weiss GM and Lockhart JW, "The impact of personalization on smartphone-based activity recognition," in AAAI Workshop on Activity Context Representation: Techniques and Languages, 2012.

[19]. Munguia Tapia E, "Using Machine Learning for Real-time Activity Recognition and Estimation of Energy Expenditure," Phd Thesis, Department of Architecture, Massachusetts Institute of Technology, Cambridge, Massachusetts, 2008.

[20]. Abdullah S, Lane ND, and Choudhury T, "Towards population scale activity recognition: A framework for handling data diversity," in Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.

[21]. Parkka J, Cluitmans L, and Ermes M, "Personalization algorithm for real-time activity recognition using PDA, wireless motion bands, and binary decision tree," Information Technology in Biomedicine, IEEE Transactions on, vol. 14, pp. 1211–1215, 2010.

[22]. Zhu X and Goldberg AB, "Introduction to semi-supervised learning," Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 3, pp. 1–130, 2009.

[23]. Guan D, Yuan W, Lee Y-K, Gavrilov A, and Lee S, "Activity Recognition Based on Semi-supervised Learning," presented at the 13th IEEE Internat Conf on Embedded and Real-Time Computing Systems and Applications, Daegu, South Corea, 2007.

[24]. Stikic M, Van Laerhoven K, and Schiele B, "Exploring Semi-Supervised and Active Learning for Activity Recognition," in 12th IEEE Internat Symp on Wearable Computers Pittsburgh, PA, 2008, pp. 81–88.

[25]. Abdallah ZS, Gaber MM, Srinivasan B, and Krishnaswamy S, "StreamAR: incremental and active learning with evolving sensory data for activity recognition," in Tools with Artificial Intelligence (ICTAI), 2012 IEEE 24th Internat Conf on, 2012, pp. 1163–1170.

[26]. Abdallah ZS, Gaber MM, Srinivasan B, and Krishnaswamy S, “Adaptive mobile activity recognition system with evolving data streams,” Neurocomputing, 2014.

[27]. Logan B, Healey J, Philipose M, Munguia Tapia E, and Intille SS, “A Long-Term Evaluation of Sensing Modalities for Activity Recognition,” in Proc of the Internat Conf on Ubiquitous Computing, Innsbruck, 2007.

[28]. Longstaff B, Reddy S, and Estrin D, “Improving Activity Classification for Health Applications on Mobile Devices using Active and Semi-Supervised Learning,” in 4th Internat Conf on Pervasive Computing Technologies for Healthcare, Munich, Germany, 2010 pp. 1 – 7

[29]. Giraud-Carrier C, “A note on the utility of incremental learning,” Ai Communications, vol. 13, pp. 215–223, 2000.

[30]. Siirtola P, Koskimäki H, and Röning J, “Personalizing human activity recognition models using incremental learning,” in The 26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2018): , Brugge, Belgium, 2018.

[31]. Intille SS, Albinali F, Mota S, Kuris B, Botana P, and Haskell WL, “Design of a wearable physical activity monitoring system using mobile phones and accelerometers,” in IEEE Engineering in Medicine and Biology Society Meeting, Boston, MA USA, 2011, pp. 3633–3639.

[32]. Albinali F, Intille S, Haskell W, and Rosenberger M, “Using wearable activity type detection to improve physical activity energy expenditure estimation,” in Proc of the 12th ACM internat conf on Ubiquitous computing, 2010, pp. 311–320.

[33]. Vapnik V, The nature of statistical learning theory. New York: Springer-Verlag, 2000.

[34]. Chang C-C and Lin C-J, “LIBSVM : A library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, vol. 2, pp. 1–27, 2011.

[35]. Wu T-F, Lin C-J, and Weng RC, “Probability estimates for multi-class classification by pairwise coupling,” Journal of Machine Learning Research, vol. 5, pp. 975–1005, 2004.

[36]. An J-L, Wang Z, and Ma Z-P, “An Incremental Learning Algorithm for Support Vector Machine,” in Proceedings of the 2nd Internat. Conference on Machine Learning and Cybernetics, Xi’an, 2003, pp. 1153–1156.

[37]. Shilton A, Palaniswami M, Ralph D, and Tsoi AC, “Incremental Training of Support Vector Machines,” IEEE Transactions on Neural Networks, vol. 16, pp. 114–131, 2005. [PubMed: 15732393]

[38]. Katagiri S and Abe S, “Incremental training of support vector machines using hyperspheres,” Pattern Recogn Letters, vol. 27, pp. 1495–1507, 2006.

[39]. Laskov P, Gehl C, Kruger S, and Muller K-R, “Incremental Support Vector Learning: Analysis, Implementation and Applications,” Journal of Machine Learning Research, vol. 7, pp. 1909–1936, 2006.

[40]. Nikitidis S, Nikolaidis N, and Pitas I, “Incremental Training of Multiclass Support Vector Machines,” in Proc of the Internat Conf on Pattern Recognition, Istanbul, Turkey, 2010, pp. 4267–4270.

[41]. Hai Y, He W, and Fan L, “An Incremental Learning Algorithm for SVM based on Voting Principle,” International Journal of Information Processing and Management, vol. 2, pp. 8–14, 2011.

[42]. Bouchachia A, Gabrys B, and Sahel Z, “Overview of Some Incremental Learning Algorithms,” Proc of FUZZ-IEEE, pp. 1–6, 2007.

[43]. Mannini A and Sabatini AM, “Machine learning methods for classifying human physical activity from on-body accelerometers,” Sensors, vol. 10, pp. 1154–1175, 2010/2/1 2010. [PubMed: 22205862]

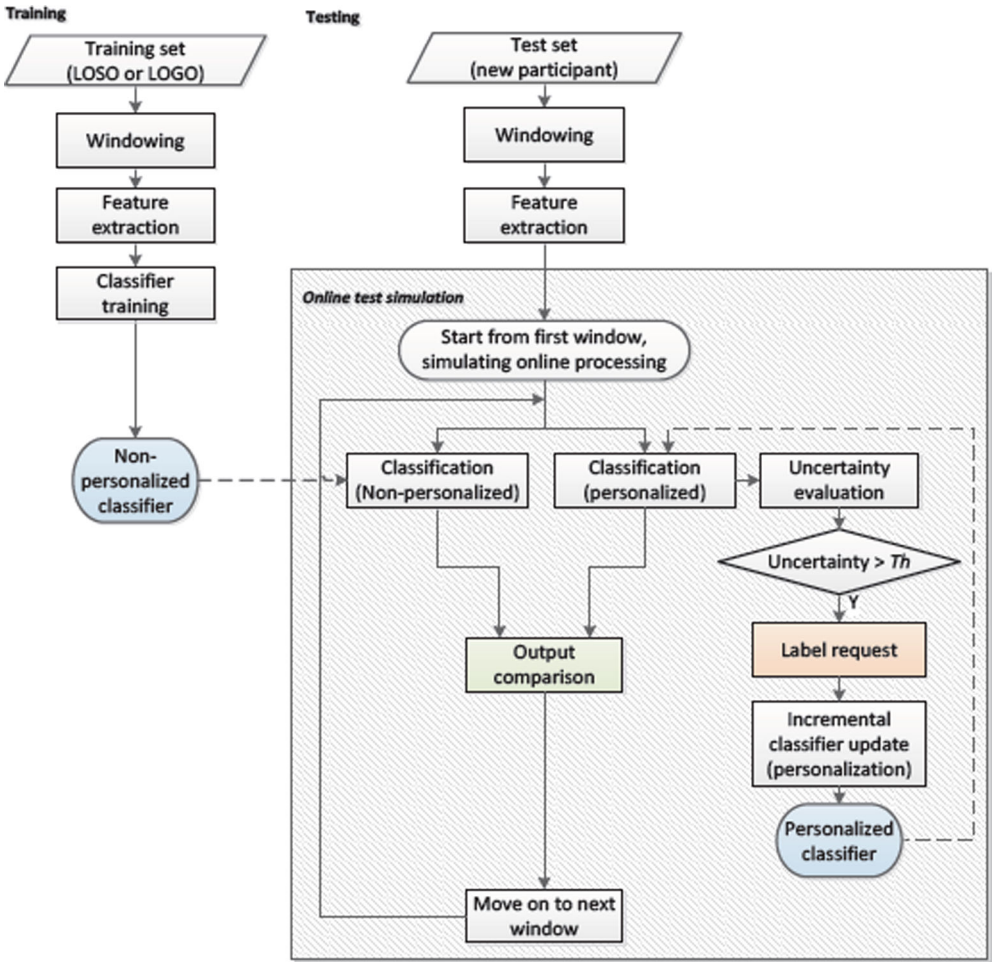


**Figure 1.**
Personalization of the classifier. Block scheme of the proposed strategy for classifier personalization testing.

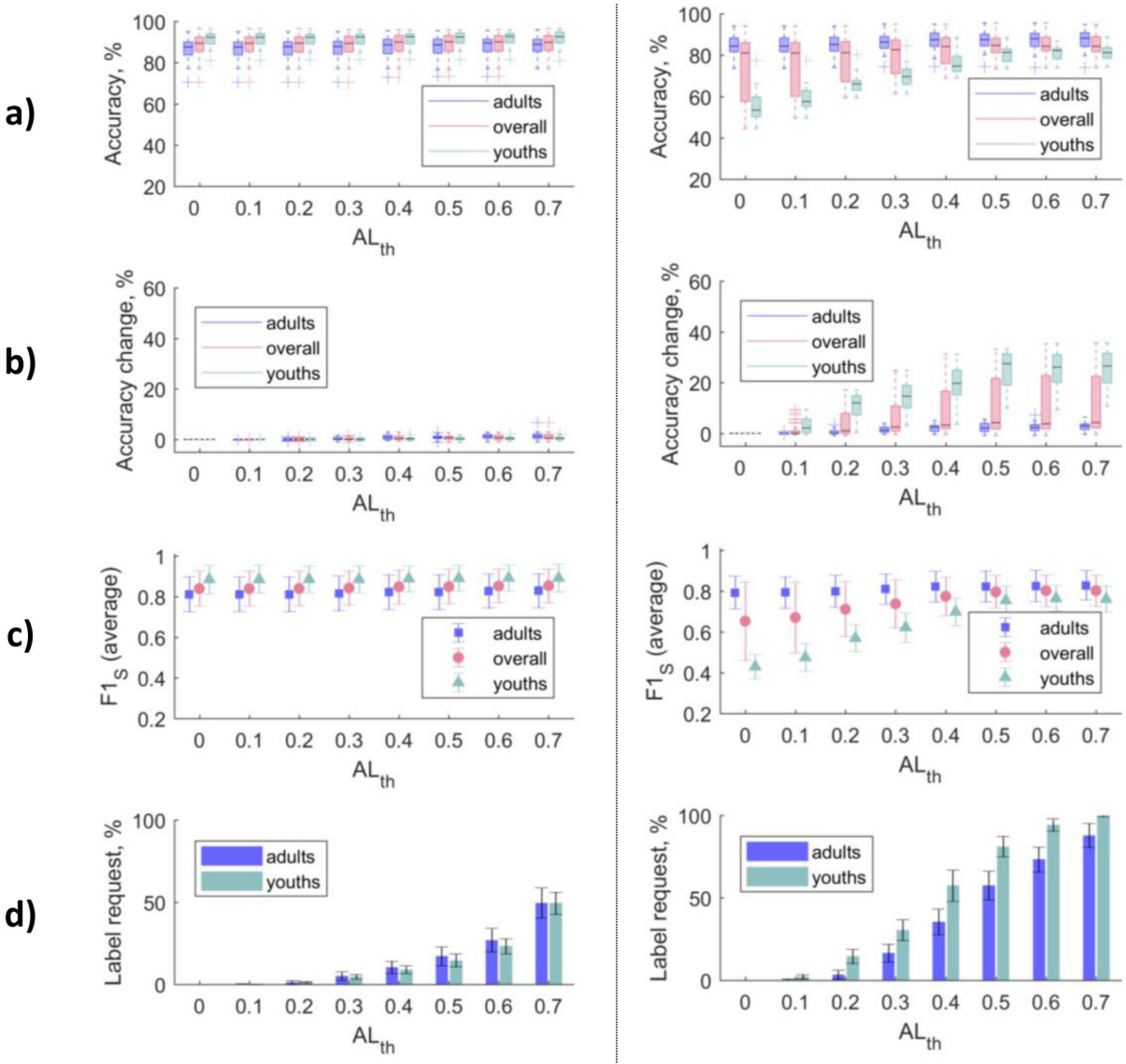


**Figure 2.**

Personalization results when varying $AL_{th}$ in the LOSO validation (left column) and LOGO validation (right column) tests: (a) boxplot of the accuracy results across the 53 participants (the line is the median, the box indicates upper and lower quartiles, the + marks are outliers); (b) boxplot of the accuracy change for each participant; (c) F1 scores ($F1_S$); (d) percentage of label requests (windows with uncertainty higher than the threshold).

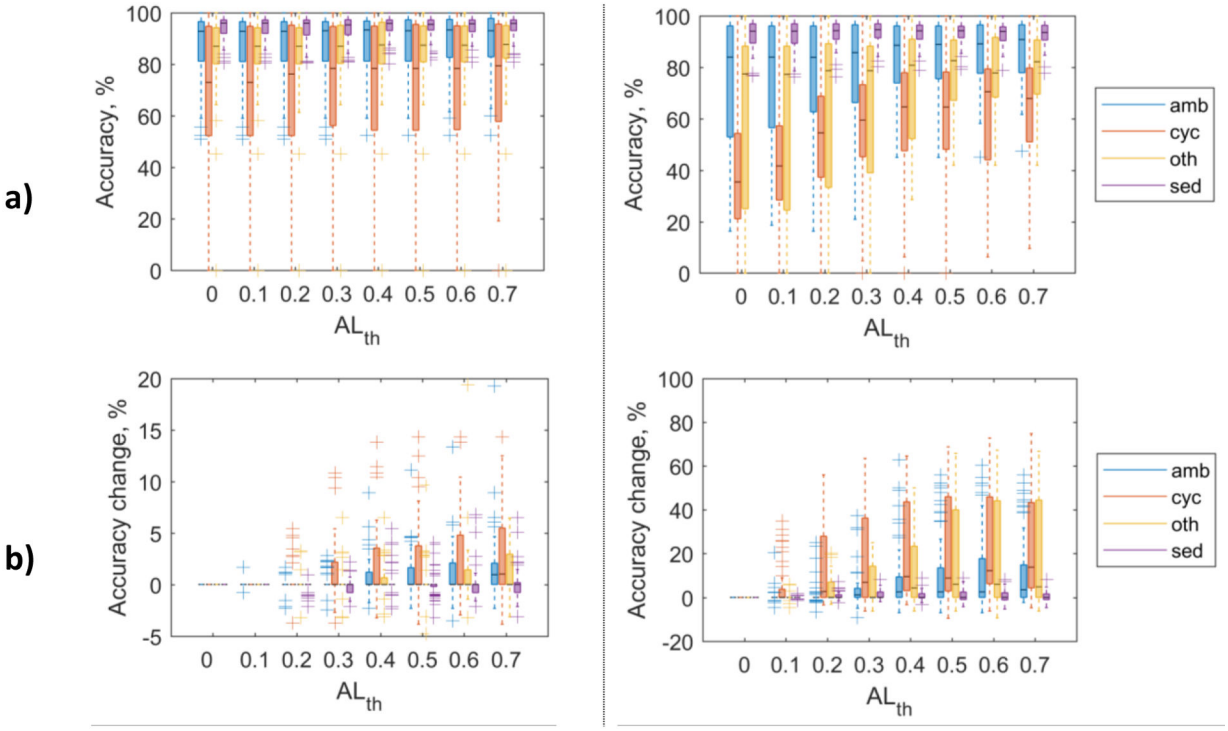


**Figure 3.**

Personalization results across the 4 classified activity groups when varying $AL_{th}$ in LOSO validation (left column) and LOGO validation (right column) tests: (a) boxplot of the accuracy results across the 53 participants (the central line is the median, the box indicates upper and lower quartiles, the + marks are outliers); (b) boxplot of the accuracy change with respect to the condition with no personalization ($AL_{th}$ = 0).

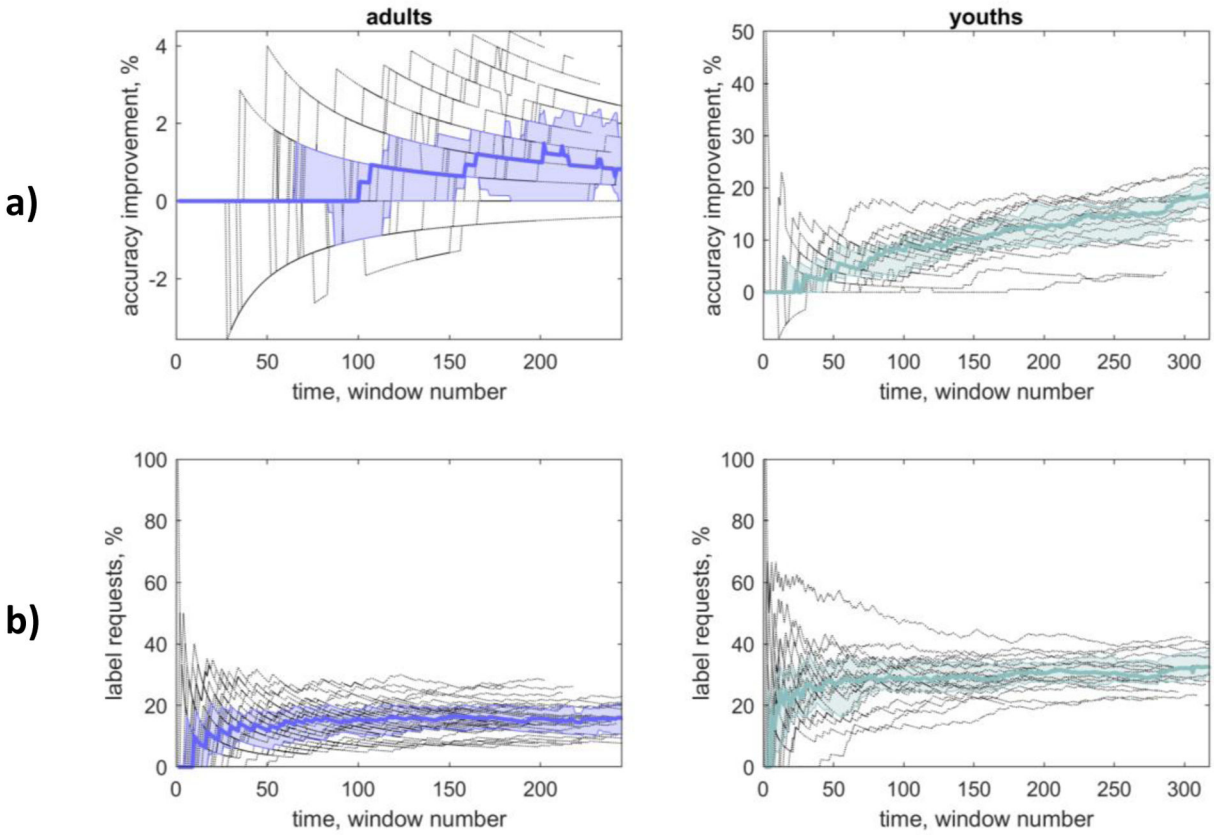


**Figure 4.**

Results for a LOGO test with $AL_{th}$ = 0.3 showing: (a) the accuracy improvement over time, as compared with the non-personalized LOGO classifier using equivalent amounts of training data and the same test data. (b) The ratio of the requested labels in the same test with respect to the total available labels so far. In all plots, the dashed lines are the results for each person in time. The solid-colored lines are the median results across all subjects, and the colored area highlights the interval between the 0.25 and 0.75 quantiles.

**TABLE I**

ACTIVITIES IN EACH OF THE AVAILABLE DATASET, GROUPED INTO FOUR CLASSES

| | Ambulation | Cycling | Other activities | Sedentary |
|---|---|---|---|---|
| **Adults datasct (A)** | • walking: natural<br>• walking: carrying-load<br>• stairs: inside and down<br>• stairs: inside and up<br>• treadmill: 3 mph 0% incline<br>• treadmill: 3 mph 6% incline<br>• treadmill: 3 mph 9% incline<br>• treadmill: 2 mph 0% incline<br>• treadmill: 4 mph 0% incline | • cycling: indoor 70rpm, 50W<br>• cycling: outdoor level<br>• cycling: outdoor uphill<br>• cycling: outdoor downhill | • painting: roller<br>• painting: brush<br>• sweeping with broom | • sitting: internet search<br>• sitting: computer typing<br>• sitting: writing<br>• sitting: reading<br>• sorting files / paperwork<br>• lying: on back<br>• lying: on left side<br>• lying: on-right-side<br>• sitting: legs straight<br>• standing still |
| **Youth dataset (Y)** | • walking: natural<br>• treadmill: 2 mph<br>• treadmill: 3–4 mph<br>• treadmill: 4.5 – 5 mph (running) | • cycling: indoor 70rpm, 50W<br>• cycling: outdoor | • basketball: dribbling<br>• basketball: passing<br>• basketball: short shots<br>• clean room<br>• soccer: dribbling<br>• soccer: passing<br>• tennis ball: fielding<br>• tennis ball: throwing-catching | • sitting: reading<br>• play computer game<br>• play on Game Boy<br>• watch TV<br>• wii: boxing<br>• wii: tennis<br>• lying: on back<br>• sitting: legs straight<br>• standing still |

**TABLE II**

RESULTS SUMMARY

| $AL_{th}$ | Label Request (%) | Change due to personalization (%) | | | Effects on classification accuracy (number of participants, %) | | | Acc(%) |
|---|---|---|---|---|---|---|---|---|
| | | *Best* | *Worst* | *Mean* | *Improved* | *No effect* | *Reduced* | |
| *LOSO* | | | | | | | | |
| 0 | - | - | - | - | - | - | - | 88.6 |
| 0.1 | 0.1 | +0.3 | −0.4 | −0.0 | 1 (2%) | 51 (96%) | 1 (2%) | 88.6 |
| 0.2 | 1.2 | +1.0 | −0.8 | +0.0 | 9 (17%) | 35 (66%) | 9 (17%) | 88.6 |
| 0.3 | 4.8 | +1.6 | −1.0 | +0.2 | 19 (36%) | 22 (42%) | 12 (22%) | 88.7 |
| 0.4 | 9.6 | +2.9 | −0.7 | +0.6 | 34 (64%) | 11 (21%) | 8 (15%) | 89.2 |
| 0.5 | 15.9 | +2.7 | −1.2 | +0.7 | 36 (68%) | 9 (17%) | 8 (15%) | 89.2 |
| 0.6 | 25.3 | +2.8 | −1.0 | +0.9 | 38 (72%) | 11 (20%) | 4 (8%) | 89.4 |
| 0.7 | 49.7 | +6.5 | −1.0 | + 1.0 | 40 (76%) | 7 (13%) | 6 (11%) | 89.5 |
| *LOGO* | | | | | | | | |
| 0 | - | - | - | - | - | - | - | 71.7 |
| 0.1 | 1.0 | +9.4 | −0.7 | +1.3 | 17 (32%) | 34 (64%) | 2 (4%) | 73.3 |
| 0.2 | 8.4 | +17.0 | −0.9 | +4.3 | 35 (66%) | 14 (26%) | 4 (8%) | 76.8 |
| 0.3 | 22.6 | +24.8 | −0.4 | +6.3 | 46 (89%) | 4 (8%) | 3 (6%) | 79.0 |
| 0.4 | 45.4 | +31.3 | −0.4 | +8.9 | 50 (94%) | 1 (2%) | 2 (4%) | 81.9 |
| 0.5 | 68.1 | +33.2 | −0.9 | +10.8 | 51 (96%) | 1 (2%) | 1 (2%) | 84.3 |
| 0.6 | 82.5 | +35.3 | −1.9 | +11.0 | 52 (98%) | 0 (0%) | 1 (2%) | 84.5 |
| 0.7 | 93.1 | +35.6 | −0.4 | + 11.1 | 49 (93%) | 3 (6%) | 1 (2%) | 84.6 |

**TABLE III**

AGGREGATED CONFUSION MATRICES AFTER LOSO AND LOGO VALIDATION

| | | **Ambulation** | **Cycling** | **Other** | **Sedentary** |
|---|---|---|---|---|---|
| *Part 1: LOSO without personalization, $AL_{th}$ = 0* | | | | | |
| | **Amb.** | **2861 (87%)** | 155 (4.7%) | 157 **(4.8%)** | 115 (3.5%) |
| | **Cyc.** | 96 (4.3%) | **1766 (78.4%)** | 30 **(1.3%)** | 360 (16%) |
| Actual label | **Oth.** | 99 (4.5%) | 43 (1.9%) | **1925 (86.6%)** | 157 (7.1%) |
| | **Sed.** | 18 (0.3%) | 168 (3.1%) | 116 (2.1%) | **5187 (94.5%)** |
| | | *Overall accuracy 88.6 %* | | | |
| *Part 2: LOSO with personalization, $AL_{th}$ = 0.4* | | | | | |
| | **Amb.** | **2887 (87.8%)** | 151 (4.6%) | 145 (4.4%) | 105 (3.2%) |
| | **Cyc.** | 88 (3.9%) | **1799 (79.9%)** | 31 (1.4%) | 334 (14.8%) |
| Actual label | **Oth.** | 93 (4.2%) | 42 (1.9%) | **1937 (87.1%)** | 152 (6.8%) |
| | **Sed.** | 19 (0.3%) | 165 (3%) | 110 (2%) | **5195 (94.6%)** |
| | | *Overall accuracy 89.2 %* | | | |
| *Part 3: LOGO without personalization, $AL_{th}$ = O* | | | | | |
| | **Amb.** | **2558 (77.8%)** | 135 (4.1%) | 213 (6.5%) | 382 (11.6%) |
| | **Cyc.** | 490 (21.8%) | **799 (35.5%)** | 307 (13.6%) | 656 (29.1%) |
| Actual label | **Oth.** | 291 (13.1%) | 30 (1.3%) | **1075 (48.3%)** | 828 (37.2%) |
| | **Sed.** | 138 (2.5%) | 195 (3.6%) | 90 (1.6%) | **5066 (92.3%)** |
| | | *Overall accuracy 71.7%* | | | |
| *Part 4: LOGO with personalization, $AL_{th}$ = 0.3* | | | | | |
| | **Amb.** | **2656 (80.8%)** | 139 (4.2%) | 171 (5.2%) | 322 (9.8%) |
| | **Cyc.** | 205 (9.1%) | **1414 (62.8%)** | 162 (7.2%) | 471 (20.9%) |
| Actual label | **Oth.** | 177 (8%) | 31 (1.4%) | **1269 (57.1%)** | 747 (33.6%) |
| | **Sed.** | 96 (1.7%) | 175 (3.2%) | 89 (1.6%) | **5129 (93.4%)** |
| | | *Overall accuracy 79.0 %* | | | |