Subject-Dependent and -Independent Human Activity Recognition with Person-Specific and -Independent Models

Sebastian Scheurer
Insight Centre for Data Analytics
School of Computer Science and Information Technology
University College Cork
Cork, Ireland
sebastian.scheurer@insight-centre.org

Salvatore Tedesco
Tyndall National Institute
University College Cork
Cork, Ireland

Kenneth N. Brown
Insight Centre for Data Analytics
School of Computer Science and Information Technology
University College Cork
Cork, Ireland

Brendan O'Flynn
CONNECT Centre for Future Networks and Communications
Tyndall National Institute
University College Cork
Cork, Ireland

ABSTRACT

The distinction between subject-dependent and subject-independent performance is ubiquitous in the Human Activity Recognition (HAR) literature. We test the hypotheses that HAR models achieve better subject-dependent performance than subject-independent performance, that a model trained with many users will achieve better subject-independent performance than one trained with a single user, and that one trained with a single user performs better for that user than one trained with this and other users by comparing four algorithms' subject-dependent and -independent performance across eight data sets using three different approaches, which we term person-independent models (PIMs), person-specific models (PSMs), and ensembles of PSMs (EPSMs). Our analysis shows that PSMs outperform PIMs by 3.5% for known users, PIMs outperform PSMs by 13.9% and ensembles of PSMs by a not significant 2.1% for unknown users, and that the performance for known users is 20.5% to 48% better than for unknown users.

iWOAR '19, September 16–17, 2019, Rostock, Germany

ACM ISBN 978-1-4503-7714-0/19/09…$15.00
https://doi.org/10.1145/3361684.3361689

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; **Ensemble methods**; *Bagging*; *Cross-validation*; Boosting; • **Human-centered computing** → *Empirical studies in ubiquitous and mobile computing*.

KEYWORDS

Human Activity Recognition; Machine Learning; Ensemble Methods; Boosting; Bagging; Inertial Sensors

ACM Reference Format:
Sebastian Scheurer, Salvatore Tedesco, Kenneth N. Brown, and Brendan O'Flynn. 2019. Subject-Dependent and -Independent Human Activity Recognition with Person-Specific and -Independent Models. In *6th international Workshop on Sensor-based Activity Recognition and Interaction (iWOAR '19), September 16–17, 2019, Rostock, Germany*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3361684.3361689

1 INTRODUCTION

Human Activity Recognition (HAR) systems are typically evaluated for their ability to generalise to either unknown users (people not represented in the HAR algorithm's training data) or to known users (people represented in the training data), with the former known as *subject-independent* and the latter as *subject-dependent* performance. The subject-independent performance can be estimated by performing a leave-one-subject-out (LOSO) cross-validation (CV) across all users in the data set, and the subject-dependent performance by performing a separate k-fold CV for each user. Which performance should be optimised when developing a HAR system depends on how it is going to be commissioned and

deployed. If commissioning a HAR system entails obtaining examples of the activities of interest from its end users—the people whose activities are to be recognised by the deployed system—then we should optimise the subject-dependent performance, which suggests that we train a personalised HAR inference model for each user. We refer to models obtained in this manner as *person-specific models* (PSMs), because they are tuned for a specific person. If, on the other hand, the system is to be deployed without prior commissioning (i.e., without being trained on data from its end users), then it must ship with a HAR model that has been pre-trained on data from a (presumably representative) sample of users. We refer to a model obtained in this manner as a *person-independent model* (PIM), because its performance is assumed to be independent of the person using it. PIMs are usually evaluated on subject-independent performance (i.e., unknown users), but it is not uncommon to see them evaluated on subject-dependent performance (known users), an approach that corresponds to a scenario where it is possible to obtain sample data from (some of) the system's end users during commissioning, but not possible to identify users (and hence the appropriate PSM) once the system has been deployed.

The distinction between subject-dependent and subject-independent performance is ubiquitous in the HAR literature, and most empirical evaluations of HAR algorithms make it clear which one was used. We intuitively hypothesise that subject-dependent performance will be better than subject-independent performance, that a PIM will outperform a PSM on subject-independent performance, and that a PSM will outperform a PIM on subject-dependent performance. Unfortunately, not many HAR papers report results for more than one combination of personalisation-generalisation approach (PIM or PSM), and subject-dependent and -independent performance, and none of them report results for all four combinations, making it impossible to verify whether these hypotheses are correct. This paper aims to narrow that gap by presenting the first empirical comparison of the subject-dependent and subject-independent performance achieved with PIM and PSM on multiple (eight) HAR data-sets, using four popular machine learning algorithms that have been used extensively and successfully in the HAR literature.

Related work

Bao and Intille [2] assess the subject-dependent performance of PSMs for recognising 20 activities of daily living (ADLs) across 20 users by training four learning algorithms on a set of semi-controlled laboratory data and evaluating them on a set of semi-naturalistic data, and the subject-independent performance of a PIM by performing a LOSO CV on the combined data from both sets. In a second experiment, they assess the subject-dependent performance of a PSM trained on three additional users' laboratory data, and the subject-independent performance of a PIM trained on five different users' laboratory data, using the three new users' semi-naturalistic data for evaluation. Unfortunately, the differences in the protocols for estimating subject-dependent and -independent performance in the first experiment means that we cannot compare them directly (the latter accuracies are 17.7% to 49.7% *higher* than the former). The second experiment, which affords a fairer comparison, directly contradicts these findings: the subject-dependent PSM accuracy (77.3%) exceeds the subject-independent PIM accuracy (73%) by 5.9%. Weiss and Lockhart [17] assess the subject-independent and -dependent performance of PIMs, and the subject-dependent performance of PSMs for recognising six ADLs across 59 users and eight learning algorithms. They report that PSMs outperform a PIM by 1.9% to 27.1% on subject-dependent accuracy, and that the subject-dependent accuracy achieved with a PIM is 11.1% to 41.1% higher than its subject-independent accuracy. These results suggest that, all else being equal, HAR methods will indeed perform better on data from known users than on data from unknown users. However, they tell us little about the size of the difference for a given personalisation-generalisation approach (PGA) or about how the trade-off between subject-dependent and -independent performance relates to the PGA.

2 METHODS

We follow the standard approach to human activity recognition comprised of data pre-processing, segmentation into windows, feature extraction from those windows, and activity inference on them based on their features [4]—where the inference step is implemented with machine learning algorithms. We estimate and compare the performance of four popular machine learning algorithms—L_2 (Ridge) regularised logistic regression, k-Nearest Neighbours (kNN), support vector machines (SVM), and a gradient boosted ensemble of decision trees (GBT)—using a set of features extracted from eight publicly available data sets, which are summarised in Table 1.

For each data set, Table 1 cites the relevant publication, lists the number of activities (act) and people (ind), and the average number of trials per activity (± standard error) and sampling frequency (Hz). We chose data sets that were acquired via wearable inertial measurement units (IMU) comprised of an acceleration and angular velocity sensor, and worn either on the chest or the wrist. Where sensors were worn on both wrists we chose the one associated with the right wrist. Unfortunately, the information about whether a user is right- or left handed is unavailable for most data-sets, making it impossible to choose the dominant wrist consistently. All data sets, except REALWORLD and SAFESENS which used a chest-worn sensor only, used a wrist-worn

Table 1: Number of (act)ivities and (ind)ividuals, trials/activity (± standard error), and sampling frequency (Hz) for each of the data-sets

	data-set	act	ind	trials/act	Hz
[13]	FUSION	7	10	90 ± 0	50
[1]	MHEALTH	11	10	38 ± 0	50
[5]	OPPORT	4	4	590 ± 258	30
[11]	PAMAP2	12	9	81 ± 8	100
[15]	REALWORLD	8	15	318 ± 42	50
[12]	SAFESENS	17	11	91 ± 13	33
[9]	SIMFALL	16	17	128 ± 8	25
[14]	UTSMOKE	7	11	859 ± 7	50

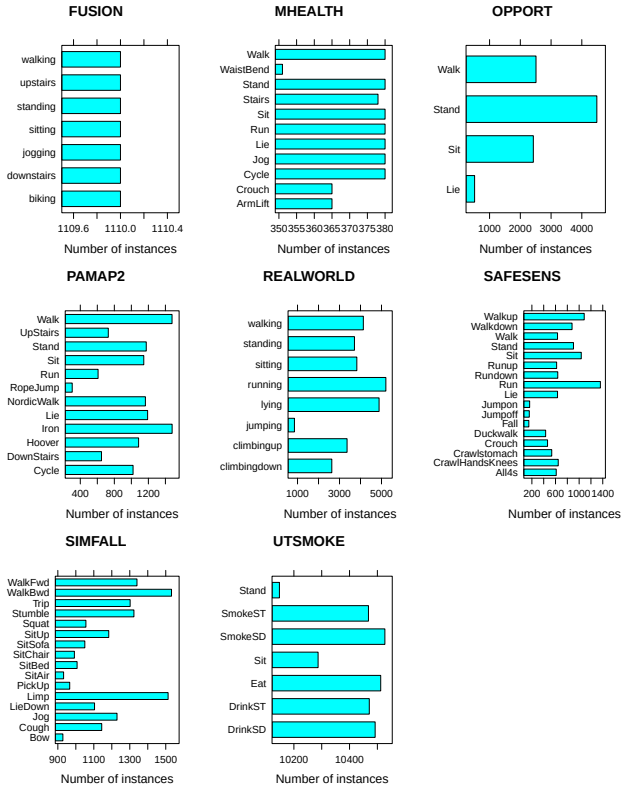

Figure 1: Number of instances per activity for each data set

sensor, and only two data-sets—PAMAP2 and SIMFALL—employed both a wrist- and a chest-worn sensor. Figure 1 illustrates how the instances—each of which corresponds to the features extracted from one window—are distributed among the activities. Note that instead of distinguishing falls from ADLs in the SIMFALL data-set, which Özdemir and Barshan [9] were able to do with Sensitivity, Specificity, and Accuracy all > 99%, we focus on the 16 ADLs shown in the figure.

The experiments were implemented in Python and parallelised via GNU parallel [16]. Analysis was carried out with *R* [10], and mixed effects models fitted with the *lme4* library [3].

We consider another personalisation-generalisation approach in addition to the person-independent (PIM) and -specific model (PSM), which we term an ensemble of PSMs (EPSM). An EPSM maintains a PSM for each known user. When an instance for a known user needs to be classified, an EPSM simply applies that user's PSM, but when an instance originates with an unknown user, it applies each user's PSM to obtain confidence scores (e.g., the estimated probability) for each activity of interest. Then the EPSM calculates each activity's mean score, and classifies the instance to the activity with the maximum mean score. To deal with the (very few) users for whom the data do not cover all the activities of interest, and whose PSMs are therefore unaware of some activities and hence unable to generate a confidence score for those activities, we assume that those activities have a probability of zero. This is not unreasonable if we accept that some people never perform certain activities (e.g., smoking, military crawling).

Pre-processing, segmentation, and feature extraction

Some data sets come with a constant timestamp for each trial—presumably introduced by storing POSIX® epoch timestamps in (sub-) millisecond resolution in Microsoft® Excel® spreadsheets. For these data-sets we generate timestamps with a fixed inter-arrival time equal to the data set's nominal sampling frequency. Then, we separate the raw data into non-overlapping *natural* trials by splitting the signal whenever the activity changes or the inter-arrival time exceeds 1.5 s. To ensure that we have at least two trials per user and activity, each of the natural trials is then split into non-overlapping batches of 15 s. Next, the body and gravity components of each trial's accelerometer signal are separated by the elliptical infinite-impulse response (IIR) low pass filter separates described by Karantonis et al. [8]. After discarding the original accelerometer data—which do not contain any information beyond that in the gravity and body components—we are left with three tri-axial signals: the gyroscope signal, the body acceleration signal, and the gravity acceleration signal. Finally, a set of time- and frequency-domain features is extracted along a sliding 3 s window with 50% (1.5 s) overlap from each trial. From the angular velocity signal and both acceleration components we extract the mean, standard deviation, skew, and kurtosis, and from the angular velocity and body acceleration signal the spectral power entropy, peak-power frequency, signal magnitude area, and the pairwise correlations between each signal's axes. This amounts to a total of 84 features that are extracted from each window.

Activity inference and evaluation

We use logistic ridge regression with $C = 0.98$, a kNN classifier with $k = 2$ and weighted voting, a SVM classifier with a radial basis function with kernel coefficient $\gamma = 0.001$ and cost penalty $C = 316$, and a GBT with a learning rate $\alpha = 0.02$ and comprised of 750 trees. The parameters for kNN, SVM, and GBT are taken from Scheurer et al. [12], who tuned them for subject-independent performance on the 17 activities in the SAFESENS data-set. The ridge parameter of $C = 0.98$ corresponds to weak regularisation, and was chosen to counteract the impact of correlated features. All features are standardised ($[x - \bar{x}]/s$) according to each feature's mean ($\bar{x}$) and standard deviation (s) in the training data. We use Cohen's Kappa (κ) to quantify the predictive performance because—unlike other performance metrics such as Sensitivity, Specificity, and Accuracy—it corrects for the probability of obtaining the observed level of agreement between the ground truth and predicted labels by chance, and because it is designed to measure predictive performance for multi-class classification.

To estimate an algorithm's subject-dependent performance, the trials are used to generate the folds in a k-fold cross validation (CV), a method we call Leave-Trials-Out (LTO) CV. LTO CV ensures that the raw data used to derive an instance in a training split are never used to derive the instances that constitute the corresponding test split, an issue that is bound to occur when working with instances derived from partially overlapping sliding windows [7], as we do here. PIM performance for known users is estimated by carrying out a k-fold LTO CV across all the users in each data-set, and PSM performance by carrying out a separate k-fold LTO CV for each user. In both cases $k = n$, where n denotes the number of people in the data-set. To estimate the subject-independent performance, we carry out a leave-m-users-out CV with $m = 1$ for EPSM and PIM, and $m = n - 1$ for PSM.

3 RESULTS, ANALYSIS, AND DISCUSSION

Figure 2 illustrates the trade-off between the performance ($\kappa \times 100$) when the user was *known*—i.e., represented in the training data—on the horizontal axis, and the performance when the user was *unknown* (not represented in the training data) on the vertical axis. In this figure, each data point corresponds to a single person (user), except in the case of person-specific models, where it corresponds to the median performance a model trained on data from the known user achieved on the other users in the data set. The symbol and colour indicate which personalisation-generalisation approach (PIM, PSM, or EPSM) was used. Table 2 summarises the results depicted in Figure 2, but using the PSM performance for all rather than, as shown in the figure, only that for the average unknown user. The table lists the mean κ (in %) ± standard error for each PGA, machine learning algorithm, data set, and sensor location.

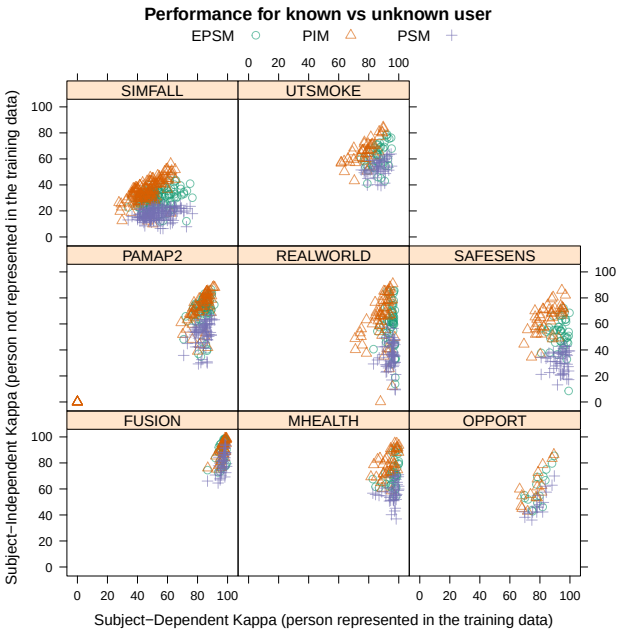

Figure 2: Subject-independent versus subject-dependent κ (%)

Analysis of the performance for known users

We can pair the performance when a person-specific model (PSM) is combined with a machine learning algorithm and applied to the data from a known person for a given data-set and sensor, to the performance when the same algorithm is combined with a person-independent model (PIM) and applied to the same data-set, sensor, and person. A paired t-test of these data yields a 95% Confidence Interval (C.I.) of 4.2 to 5.3 percentage points (hereafter, points) for the difference between the κ achieved with PSMs and that achieved with PIM, with a mean difference of 4.8 points ($t_{443} = 16.7$, $P < 2.2 \times 10^{-16}$), suggesting that we can be 95% confident that a PSM outperforms a PIM on data from known users by 4.2 to 5.3 points. However, it is unlikely that the t-test's underlying assumption of identically and independently distributed (IID) data is met, because the difference in the subject-dependent performance between PIM and PSM might depend not only on the data-set—which is expected due to the different activities of interest, and evident in Figure 2—but also on the learning algorithm.

A more appropriate tool for analysing data, such as these, that are not IID is the *linear mixed effects model* (LMM). A LMM extends linear regression with so-called *random effects* which allow us to impose structure on the residuals. We can,

Table 2: κ (%) ± standard error when learning algorithms (mla) are combined with a person-independent model (PIM), a person-specific model (PSM), or an ensemble of PSMs (EPSM), and tested on known or unknown users

			known		unknown		
dataset	sensor	mla	PSM/EPSM	PIM	PSM	EPSM	PIM
FUSION	wrist	gbt	97.7 ± 0.4	97.9 ± 0.4	82.6 ± 2.6	90.4 ± 2.6	92.6 ± 2.1
		knn	94.2 ± 1.0	94.1 ± 0.9	76.6 ± 1.4	87.5 ± 2.8	85.9 ± 2.0
		logreg	97.4 ± 0.4	96.7 ± 0.6	81.4 ± 2.0	89.5 ± 2.8	91.9 ± 2.1
		svm	97.8 ± 0.4	97.7 ± 0.3	81.8 ± 2.0	90.3 ± 2.7	90.9 ± 2.1
MHEALTH	wrist	gbt	97.1 ± 1.2	96.8 ± 1.0	59.1 ± 1.7	72.3 ± 3.6	82.3 ± 3.4
		knn	93.7 ± 1.4	91.3 ± 1.8	55.6 ± 2.0	71.4 ± 2.6	76.1 ± 3.1
		logreg	95.7 ± 1.4	91.9 ± 1.5	54.6 ± 2.3	70.0 ± 2.9	78.9 ± 3.3
		svm	96.8 ± 1.0	94.4 ± 1.3	58.9 ± 2.3	72.0 ± 3.9	82.0 ± 2.6
OPPORT	wrist	gbt	83.3 ± 2.4	81.7 ± 2.6	58.7 ± 4.3	66.9 ± 8.1	68.8 ± 7.1
		knn	74.8 ± 2.7	71.3 ± 2.5	40.8 ± 1.7	54.5 ± 5.4	51.4 ± 4.2
		logreg	76.7 ± 3.0	72.6 ± 3.5	46.6 ± 3.9	56.7 ± 6.7	59.9 ± 7.0
		svm	81.0 ± 2.4	81.5 ± 2.3	46.2 ± 1.7	61.7 ± 6.6	65.4 ± 6.7
PAMAP2	chest	gbt	87.7 ± 0.7	77.7 ± 9.7	57.9 ± 3.5	72.6 ± 4.3	69.4 ± 9.3
		knn	78.5 ± 1.2	67.2 ± 8.4	51.0 ± 3.2	67.7 ± 3.2	56.5 ± 7.4
		logreg	85.4 ± 0.9	73.5 ± 9.2	50.2 ± 3.4	69.3 ± 4.9	64.2 ± 8.7
		svm	85.1 ± 0.8	76.4 ± 9.6	51.3 ± 3.7	69.4 ± 5.1	65.7 ± 9.1
PAMAP2	wrist	gbt	85.9 ± 0.8	76.9 ± 9.7	57.2 ± 2.7	71.5 ± 2.9	70.2 ± 9.1
		knn	78.9 ± 1.6	68.7 ± 8.7	49.9 ± 4.6	68.1 ± 3.9	57.9 ± 8.1
		logreg	83.5 ± 1.3	72.8 ± 9.2	53.7 ± 4.4	68.8 ± 4.9	66.3 ± 9.0
		svm	83.3 ± 1.3	75.3 ± 9.5	47.4 ± 3.8	68.2 ± 4.5	65.0 ± 9.3
REALWORLD	chest	gbt	96.1 ± 0.4	93.3 ± 0.6	40.2 ± 3.5	62.4 ± 3.9	71.7 ± 4.4
		knn	91.3 ± 1.0	85.1 ± 1.5	40.2 ± 3.1	61.8 ± 3.7	59.3 ± 3.4
		logreg	95.4 ± 0.5	83.8 ± 1.8	32.4 ± 3.4	57.2 ± 4.2	60.7 ± 5.8
		svm	95.5 ± 0.4	92.0 ± 0.7	31.9 ± 3.8	54.9 ± 4.5	62.4 ± 5.1
SAFESENS	chest	gbt	97.0 ± 0.8	93.5 ± 0.8	29.3 ± 2.9	47.9 ± 5.0	68.3 ± 3.5
		knn	87.8 ± 1.5	81.1 ± 1.7	31.7 ± 3.4	54.7 ± 3.2	55.7 ± 3.5
		logreg	93.1 ± 1.0	78.4 ± 1.5	27.2 ± 2.7	54.0 ± 2.7	64.1 ± 3.0
		svm	95.2 ± 0.8	87.9 ± 1.1	30.8 ± 3.4	51.9 ± 2.6	66.9 ± 2.7
SIMFALL	chest	gbt	65.9 ± 1.3	57.1 ± 1.2	19.8 ± 0.8	33.5 ± 1.5	44.0 ± 1.7
		knn	50.0 ± 1.3	44.9 ± 0.9	20.4 ± 0.6	33.3 ± 1.0	30.3 ± 0.7
		logreg	52.3 ± 1.2	38.0 ± 0.9	14.8 ± 0.7	29.1 ± 1.0	34.5 ± 1.1
		svm	49.5 ± 1.6	49.8 ± 0.8	14.3 ± 0.6	25.9 ± 1.0	38.2 ± 1.4
SIMFALL	wrist	gbt	62.4 ± 1.5	55.6 ± 1.5	20.2 ± 0.8	32.6 ± 2.2	40.4 ± 2.3
		knn	48.8 ± 1.2	44.6 ± 1.2	20.1 ± 0.7	31.8 ± 1.6	29.2 ± 1.5
		logreg	49.6 ± 1.3	37.1 ± 1.4	17.3 ± 0.7	27.9 ± 1.7	32.7 ± 2.1
		svm	45.9 ± 1.5	48.1 ± 1.3	14.4 ± 0.7	27.1 ± 1.5	36.0 ± 2.3
UTSMOKE	wrist	gbt	90.8 ± 0.9	81.0 ± 1.5	55.7 ± 1.2	65.3 ± 3.3	68.8 ± 2.9
		knn	81.2 ± 1.2	76.2 ± 1.3	51.8 ± 1.5	60.7 ± 2.8	61.6 ± 2.4
		logreg	84.1 ± 1.2	69.0 ± 2.0	51.0 ± 1.0	59.4 ± 2.5	63.2 ± 2.5
		svm	89.1 ± 0.9	83.8 ± 1.3	53.7 ± 1.6	63.6 ± 2.9	69.2 ± 2.7

for example, specify that the performances within data-sets are correlated, or even that the difference in performance between PGAs varies depending on the data-set. The random effects are assumed to add up to zero, and hence the fixed effects (which are analogous to linear regression coefficients) can be estimated via (restricted) maximum likelihood. Unfortunately, there is no consensus on how to obtain P-values for LMM coefficients, but it is possible to obtain C.I.s via likelihood profiling, bootstrap sampling, or by making assumptions about the likelihood function's shape in which case a Wald test can be used. For a detailed treatment of LMMs we refer interested readers to Gelman and Hill [6].

A LMM that models the subject-dependent performance ($\ln[\kappa + 1]$, to be precise) as a combination of (fixed) effects attributable to the machine learning algorithm and personalisation-generalisation approach—either PIM or PSM, since subject-dependent EPSM performance is identical to that of its constituent PSMs—and a random effect to control for the variation of the PGA effect between data sets, explains the observed variation in the response with a residual standard deviation of 6.3 points. This model reveals that the (random) effect of applying PSM to a data-set, which varies with a standard deviation of 1.9 points between data-sets, is moderately inversely correlated (−0.53) with PIM performance, which varies with a standard deviation of 9.3 points, on the same data-set. This confirms the intuition that a PSM likely confers less advantage when applied to a data-set on which a PIM performs well. The restricted maximum likelihood (REML) estimates of the fixed effects indicate that GBT—with an estimated κ of 84.6% and a 95% (bootstrap) C.I. of 71.9% to 97.1% when used as a PIM—outperforms SVM by 2.9% (1.7% to 4.1%), logistic regression by 5.4% (4.2% to 6.6%), and kNN by 5.5% (4.1% to 6.7%), regardless of the PGA. They further show that PSMs outperform the corresponding PIM by 3.5% (1.7% to 5.2%) when evaluated on known users.

Analysis of the performance for unknown users

A paired t-test for the difference between PIM and EPSM performance for unknown users yields a 95% C.I. of 4.4 to 5.8 points with a mean difference of 5.1 points ($t_{443} = 14.7$, $P < 2.2 \times 10^{-16}$), indicating that PIM significantly outperforms EPSM. A paired t-test for the difference between subject-dependent PIM performance and median PSM performance for unknown users yields a 95% C.I. of 19.1 to 21 points with a mean difference of 20 points ($t_{443} = 41$, $P < 2.2 \times 10^{-16}$). Finally, a paired t-test for the difference between EPSM and median PSM (subject-independent) performance yields a 95% C.I. of 14.2 to 15.6 points with a mean difference of 14.9 points ($t_{443} = 41.3$, $P < 2.2 \times 10^{-16}$). These results indicate that a PIM outperforms an EPSM by 5.1 points and the average PSM by 20 points, and that an EPSM outperforms its average constituent PSM by 14.9 points when evaluated on data from a user who was not represented in the training data.

The same LMM fitted to the subject-independent performance ($\ln[\kappa - 1]$) explains the observed variation in the response with a residual standard deviation of 8.6 points. This model reveals that the (random) effect of applying EPSM to a data-set, which varies with a standard deviation of 3.2 points between data-sets, weakly correlates (0.21) with PIM performance (EPSM is likely to perform better on data sets on which PIM performs better, too), that the PSM effect, which varies with a standard deviation of 6.4 points, correlates weakly (0.31) with PIM performance, which varies with a standard deviation of 9.7 points, and that the PSM and EPSM effects correlate strongly (0.78) with each other (PSM and EPSM tend to perform better on the same data sets). The REML estimates of the fixed effects indicate that GBT—with an estimated κ of 67.5%, and a 95% C.I. of 56.1% to 78.9% when used as a PIM—outperforms kNN by 1.3% (0.7% to 1.9%), SVM and logistic regression by 3.4% (with respective 95% C.I.s of 2.7% to 4% and 2.8% to 4.1%), regardless of the PGA. The estimates further show that although a PIM outperforms the corresponding PSMs by a clear 13.9% (9.1% to 19.3%), it only narrowly beats an EPSM by a not significant 2.1% (−0.3% to 5%).

Discussion

These results suggest that the best subject-dependent performance is achieved with PSMs, and the best subject-independent performance with PIMs. Hence, in order to optimise both subject-dependent and -independent performance, we should use PSMs for known users and a PIM for unknown users wherever possible. If we use a PIM, rather than a PSM, for known users we forego an expected improvement of 3.5% in the subject-dependent performance, and if we use a PSM (rather than a PIM) for unknown users we forego an expected improvement of 13.9% in the subject-independent performance. Perhaps surprisingly, we found that although an EPSM—whose subject-dependent performance is identical to that of its constituent PSMs—does perform 2.1% worse than a PIM for unknown users, that estimate comes with a 95% C.I. of –0.3% to 5%, according to which we do not have enough evidence to reject the (null) hypothesis that there is no real difference in the subject-independent performance between PIM and EPSM.

With respect to the difference between the subject-dependent and -independent performance, our findings imply that a PIM performs 20.5% to 26.1% better for known users than for unknown users, an EPSM 27.5% to 33.3% better, and a PSM 42% to 48.5% better. In all cases the difference is smallest with kNN, followed by logistic regression with differences of 23.1%, 30.2%, and 45.1%, GBT with differences of 25.4%,

32.6%, and 47.8%, and SVM with differences of 26.1%, 33.3%, and 48.5%, respectively.

4 CONCLUSION

We compared the subject-dependent and -independent performance of person-independent models (PIMs) and person-specific models (PSMs), and ensembles of PSMs (EPSMs) when used with four popular HAR algorithms across eight publicly available HAR data-sets. Analysis with mixed effects models showed that GBT outperforms the other algorithms on both subject-dependent and -independent performance, that PSMs outperform a PIM by 3.5% when evaluated on known users, and that a PIM outperforms a PSM by 13.9% when evaluated on unknown users. The analysis further showed that although PIMs outperform EPSMs on unknown users, according to the 95% C.I. the 2.1% difference is not significant. We estimated that the difference between subject-dependent performance and subject-independent performance ranges from 20.5% to 26.1% with PIMs, 27.5% to 33.3% with EPSMs, and from 42% to 48.5% with PSMs.

ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland and the European Regional Development Fund under grant number SFI/12/RC/2289 and grant number 13/RC/2077-CONNECT, as well as the European funded project SAFESENS under the ENIAC program in association with Enterprise Ireland under grant number IR20140024.

REFERENCES

[1] O. Baños, R. Garcia, J. A. Holgado-Terriza, M. Damas, H. Pomares, I. Rojas, A. Saez, and C. Villalonga. 2014. mHealthDroid: a novel framework for agile development of mobile health applications. In *Int. Workshop on Ambient Assisted Living*. Springer. DOI: 10.1007/978-3-319-13105-4_14.

[2] L. Bao and S. S. Intille. 2004. Activity recognition from user-annotated acceleration data. In *Int. Conf. on Pervasive Computing*. DOI: 10.1007/978-3-540-24646-6_1.

[3] D. Bates, M. Mächler, B. Bolker, and S. Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1. DOI: 10.18637/jss.v067.i01.

[4] A. Bulling, U. Blanke, and B. Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *Comput. Surv.*, 46, 3, (January 2014). DOI: 10.1145/2499621.

[5] R. Chavarriaga, H. Sagha, A. Calatroni, S. T. Digumarti, G. Tröster, J. d. R. Millán, and D. Roggen. 2013. The opportunity challenge: a benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*, 34, 15. Smart Approaches for Human Action Recognition. DOI: 10.1016/j.patrec.2012.12.014.

[6] A. Gelman and J. Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

[7] A. Jordao, A. C. Nazare Jr., J. Sena, and W. R. Schwartz. 2019. Human activity recognition based on wearable sensor data: A standardization of the state-of-the-art. *CoRR*, (February 2019). arXiv: 1806.05226v3 [cs.CV].

[8] D. M. Karantonis, M. R. Narayanan, M. Mathie, N. H. Lovell, and B. G. Celler. 2006. Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. *Trans. on Information Technology in Biomedicine*, 10, 1, (January 2006). DOI: 10.1109/TITB.2005.856864.

[9] A. T. Özdemir and B. Barshan. 2014. Detecting falls with wearable sensors using machine learning techniques. *Sensors*, 14, 6, (June 2014). DOI: 10.3390/s140610691.

[10] R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Version 3.3.3. R Foundation for Statistical Computing. Vienna, Austria.

[11] A. Reiss and D. Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *Int. Symposium on Wearable Computers*. IEEE, (June 2012). DOI: 10.1109/ISWC.2012.13.

[12] S. Scheurer, S. Tedesco, K. N. Brown, and B. O'Flynn. 2017. Human activity recognition for emergency first responders via body-worn inertial sensors. In *Int. Conf. on Wearable and Implantable Body Sensor Networks*. IEEE, (May 2017). DOI: 10.1109/BSN.2017.7935994.

[13] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. M. Havinga. 2014. Fusion of smartphone motion sensors for physical activity recognition. *Sensors*, 14, 6. DOI: 10.3390/s140610146.

[14] M. Shoaib, H. Scholten, P. J. M. Havinga, and O. D. Incel. 2016. A hierarchical lazy smoking detection algorithm using smartwatch sensors. In *Int. Conf. on e-Health Networking, Applications and Services*. (September 2016). DOI: 10.1109/HealthCom.2016.7749439.

[15] T. Sztyler and H. Stuckenschmidt. 2016. On-body localization of wearable devices: an investigation of position-aware activity recognition. In *Int. Conf. on Pervasive Computing and Communications*. IEEE, (March 2016). DOI: 10.1109/PERCOM.2016.7456521.

[16] O. Tange. 2011. GNU parallel—the command-line power tool. *;login: The USENIX Magazine*, 36, 1, (February 2011).

[17] G. M. Weiss and J. W. Lockhart. 2012. The impact of personalization on smartphone-based activity recognition. In *Conf. on Artificial Intelligence*. Workshop on Activity Context Representation: Techniques and Languages. AAAI.