Evaluating the transferability of personalised exercise recognition models.

WIJEKOON, A. and WIRATUNGA, N.

2020

Evaluating the Transferability of Personalised Exercise Recognition Models*

Anjana Wijekoon$^{1[0000-0003-3848-3100]}$ and Nirmalie Wiratunga$^{1[0000-0003-4040-2496]}$

Robert Gordon University, Aberdeen AB10 7GJ, Scotland, UK
{a.wijekoon, n.wiratunga}@rgu.ac.uk

Abstract. Exercise Recognition (ExR) is relevant in many high impact domains, from health care to recreational activities to sports sciences. Like Human Activity Recognition (HAR), ExR faces many challenges when deployed in the real-world. For instance, typical lab performances of Machine Learning models, are hard to replicate, due to differences in personal nuances, traits and ambulatory rhythms. Thus effective transferability of a trained ExR model, depends on its ability to adapt and personalise to new users or user groups. This calls for new experimental design strategies that are also person-aware, and able to organise train and test data differently from standard ML practice. Specifically, we look at person-agnostic and person-aware methods of train-test data creation, and compare them to identify best practices on a comparative study of personalised ExR model transfer. Our findings show that ExR when compared to results with other HAR tasks, to be a far more challenging personalisation problem and also confirms the utility of metric learning algorithms for personalised model transfer.

Keywords: Exercise Recognition · Transferability · Personalisation · Performance Evaluation

1 Introduction

Exercise Recognition (ExR) is an ongoing Machine Learning (ML) research challenge with many practical applications such as self-management of musculoskeletal pain, weight training, orthopaedic rehabilitation and strength and balance improvement of pre-frail adults. Research in ExR falls under Human Activity Recognition (HAR) research, which has broader applications in gait recognition, fall detection and activity recognition for fitness applications, to name a few.

Fitness applications that adopt ExR as an integral component, face many challenges at deployment, compared with other conventional Machine Learning (ML) or Deep Learning (DL) applications such as image recognition or text classification. For instance lack of transferability of learned ML models is one of

* This work is part funded by SELFBACK which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 689043.

the main challenges that is present in many forms such as; the transferability to new sensor modalities, to new activities or to new user groups. With new sensor modalities, both heterogeneity of sensor data and differences in sensor configurations must be addressed. Transferability to new activity classes is generally addressed as open-ended HAR, where either a knowledge-intensive method is used with a corpus to learn heuristics that can cover all possible activities classes to be expected in the future [5] or; in contrast a knowledge-light method learns a feature space that is expected to adapt to new activity classes [13]. Lastly, when deploying a generic fitness application, developers are unaware of the target user group; and are unable to make recommendations or adapt to alternative sensors, and/or exercise or physical activities. Here transferability to a new user group, in addition to variations in sensor modalities, must also consider common factors applicable to the group needs.

In this paper, we focus on ExR applications; their transferability to different user groups and importantly how to design comparative studies that are informed by ownership of data (i.e. data that is generated by a specific person). Naturally, people incorporate many personal nuances when performing exercises. In practice, these personal traits are captured by sensors. If the ExR algorithm is unaware of the specific person it may find it challenging to map sensor readings to a specific exercise. In this paper we show that adopting the correct evaluation method is crucial to understanding the capabilities of an ExR algorithm amongst a diverse group of users.

We explore person-aware evaluation methods using the HAR personalisation algorithm MN^p [13] that was inspired by Metric-Learning and Meta-Learning ideas in the HAR (physical activity) domain. MN^p achieves personalisation without requiring test user data and learns a feature space that is transferable to a wider range of users. We expect ExR to be a harder personalised model transfer challenge, compared to models for recognising physical activities. Lets consider a typical ambulatory physical activity such as walking, where gait and personal nuances easily influence walking cycles; contrast that to an exercise, such as a pelvic tilt or a knee roll (see Figure 4a) where its harder to isolate and capture personal nuances. Our evaluation shows that MN^p can be applied to ExR and is transferable to new users not seen during training.

Rest of the paper is organised as follows. In Section 2 we explore related literature in HAR and ExR domains. Next in Section 3, we present methods that are explored in this paper; followed by an analysis of results with alternative evaluation strategies in Section 4. Finally conclusions and future directions appear in Section 5.

2 Related Work

Research in Exercise Recognition (ExR) covers a wide variety of application areas such as weight training [10, 6], rehabilitation [2] and callisthenics and gym exercises [8, 15]. ExR like HAR, is a multi-class classification problem where classes are unique exercises that are captured by a stream of sensor data. Many algorithms have been explored in literature such as k-NN [8, 14], Decision Trees and Random Forest [10, 15] and CNN and LSTM [2, 12].

Personalising such algorithms is intuitively desirable for ExR as personal nuances such as gait, posture and rhythm are known factors that are used by human experts when analysing exercise performance and adherence in the real-world. Although this remains largely unexplored for ExR; there is useful work in Human Activity Recognition (HAR), where early research has looked at user-dependent modelling with access to large quantities of labelled end-user data [9, 1, 7]. Follow on work attempts to reduce this human burden, by adopting semi-supervised learning methods [3, 4] that require some model re-training after deployment.

Recent advances in few-shot learning and meta-learning with Matching Networks (MN) [11] and Personalised MN (MN^p) for HAR [13] has addressed the short comings of previous methods by only using few data instances from end-user as well as learning embeddings that are largely transferable to new activities without needing model re-training after deployment. It has also outperformed its non-personalised counterpart in the tasks of pose detection and HAR [13]. Importantly in this paper we investigate, if MN^p is transferable to ExR, which is arguably a more challenging personalised learning problem.

3 Methods

Given sensor data streams recorded while performing exercises, for supervised ExR, data instances are extracted using the sliding window method applied to each of the streams. Typical sensors include inertial sensors, depth cameras and pressure mats. More formally, given a set of data instances, $\mathcal{X}$, ExR involves learning a feature space where the mapping from each instance, x, to an exercise class, y, where y is from the set of exercise classes, $\mathcal{L}$. Accordingly, each sensor-based data instance is a data and class label pair, (x, y), where $y \in \mathcal{L}$.

$$\mathcal{X} = \{(x, y) \mid y \in \mathcal{L}\} \tag{1}$$

In comparison to computer vision or text datasets, each data instance in $\mathcal{X}$ belongs to a person, p. Given the set of data instances obtained from person p is $\mathcal{X}^p$, relationship of $\mathcal{X}^p$ and $\mathcal{X}$ formalised as in Eqation 2. As before all data instances in $\mathcal{X}^p$ will belong to a class in $\mathcal{L}$ except special instances like open-ended HAR where the class set is not fully specified at training time.

$$\mathcal{X} = \{\mathcal{X}^p \mid p \in \mathcal{P}\} \text{ where } \mathcal{X}^p = \{(x, y) \mid y \in \mathcal{L}\} \tag{2}$$

Training and testing methodologies can adopt one of two approaches; *person-agnostic* where an algorithm is trained and tested with the same user group; and *person-aware*, where an algorithm is trained and tested on different user groups. Both maintain disjointed sets of data instances in train and test; but the latter also preserves disjoint persons by preserving the person-to-data relationship during model training and testing.

3.1 Person-agnostic Evaluation

Person agnostic evaluations can be applied as a repeated hold-out (R-HO) or a cross-fold (CF) validation methodologies. In either case, the person parameter of each data instance is discarded when creating hold-out sets, or folds. This means that a person's data can be split between train and test sets. A percentage, λ, of all data instance in $\mathcal{X}$, is used as the set of test data instances, $\mathcal{X}_{test}$, and the rest as the set of train data instances $\mathcal{X}_{train}$.

$$\begin{gathered}\mathcal{X} = \{\mathcal{X}_{train}, \mathcal{X}_{test}\} \\ |\mathcal{X}_{train}| \approx (1-\lambda) \times |\mathcal{X}| \text{ and } |\mathcal{X}_{test}| \approx \lambda \times |\mathcal{X}|\end{gathered} \tag{3}$$

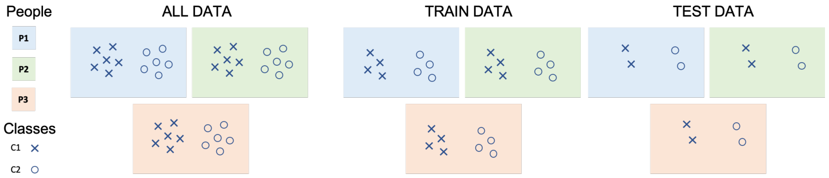


Fig. 1: Person-agnostic train/test split

With R-HO, a λ percentage of the data instances are randomly selected (without replacement) to form the test set and the remainder forming the train set. This is repeated for multiple iterations. With CF, first, the dataset is divided into a number of folds where each fold contains λ percentage of data, and at each iteration, one fold is selected as the test set and the rest of the folds as the training set. Both methods create train and test sets that share the same population $\mathcal{P}$ (see Figure 1). Unlike with R-HO method, CF guarantees that each data instance appears once in the test set.

A person-agnostic methodology for evaluation, trains and tests on the same population. Accordingly, these methodologies are not designed to evaluate the robustness of an algorithm on a different population following typical deployment. However they provide an "upper-bound" performance of a ML algorithm.

3.2 Person-aware Evaluation

A Person-aware evaluation can be performed as a repeated Persons-Hold-Out (R-PHO) or a Leave-One-Person-Out (LOPO) methodology. With R-PHO, a percentage, μ, of the user population is selected as the test user set, rest forming the train user set; and this is repeated for multiple iterations. With LOPO methodology, instances from a single user is put aside, to form a singleton test user group, and the rest of the users form the training user group. The train and test set formation with the test user group, $\mathcal{P}_{test}$, and the train user group, $\mathcal{P}_{train}$

can be formalised as follows:

$$\begin{gathered}\mathcal{X} = \{\mathcal{X}_{train}, \mathcal{X}_{test}\} \\ \mathcal{X}_{test} = \{\mathcal{X}^p \mid p \in \mathcal{P}_{test}\} \text{ and } \mathcal{X}_{train} = \{\mathcal{X}^p \mid p \in \mathcal{P}_{train}\} \\ \mathcal{P} = \{\mathcal{P}_{train}, \mathcal{P}_{test}\}\end{gathered} \tag{4}$$

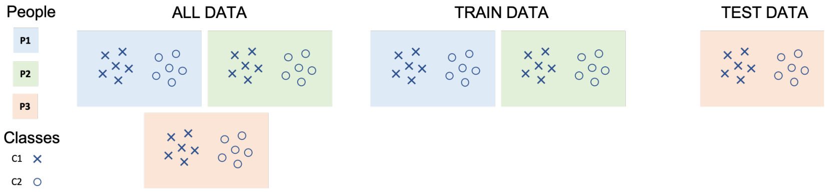


Fig. 2: Person-aware train/test split

LOPO ensures that each user in the population $\mathcal{P}$ is included in the test set in one of the trials (similar to the person-agnostic CF method), but LOPO also ensures disjointedness in selected persons (see Figure 2). We propose that performance measures obtained with a person-aware methodology should be used as the "lower-bound", likely performance of a ML algorithm after deployment.

3.3 Personalised Matching Networks

The goal of personalisation is to learn a feature space that can dynamically adapt to different test user groups. With reference to Sections 3.1 and 3.2, the aim here is to find algorithms that outperform the "lower-bound" set by a non-personalised algorithm when evaluated by a person-aware methodology. For this purpose we explore the Personalised Matching Networks (MN^p), which has been successfully used for personalising HAR algorithms.

MN^p is inspired by Metric Learning and Meta-Learning paradigms where the classification task is learning to match a test instance, x, to one instance from a set of representatives. The set of representative instances, S is chosen from the same user (i.e. from. $\mathcal{X}^p$) ensuring all classes are represented. An instance in, S, is a data instance, (x, y), and for each class up to, k, representatives are selected from, $\mathcal{X}^p$, as in Equation 5.

$$S = \{(x, y)|x \in \mathcal{X}^p, y \in \mathcal{L}\} \text{ where } |S| = k \times |\mathcal{L}| \tag{5}$$

We denote the training data set obtained for person, p's data as $\mathcal{X}^p_{tr}$. An instance in, $\mathcal{X}^p_{tr}$, consists of a query and support set pairs, (q_i, S_i), where, q_i, is a sensor data and class label pair, (x_i, y_i), (similar to a conventional supervised learning training data instance). The complete training data set, $\mathcal{X}_{tr}$, is the collection of all, $\mathcal{X}^p_{tr}$, for the train user group $\mathcal{P}_{tr}$.

$$\begin{gathered}\mathcal{X}^p_{tr} = \{(q,S) \mid x \in \mathcal{X}^p, y \in \mathcal{L}\} \text{ where } q=(x,y), y \in \mathcal{L} \\ \mathcal{X}_{tr} = \{\mathcal{X}^p_{tr} \mid p \in \mathcal{P}_{tr}\}\end{gathered} \tag{6}$$

During MN^p training, a model learns a feature space where data instances from different users are successfully transformed and mapped to class labels. Here training can be viewed as a parameterised (Θ) end-to-end learning of a distance / similarity function, using a non-parametric attention-based kernel to compute the objective matching function (see architecture in Figure 3). At testing, the MN^p algorithm predicts the label $\hat{y}$ for a query instance $\hat{x}$ with respect to its support set $\hat{S}$ from the same user.

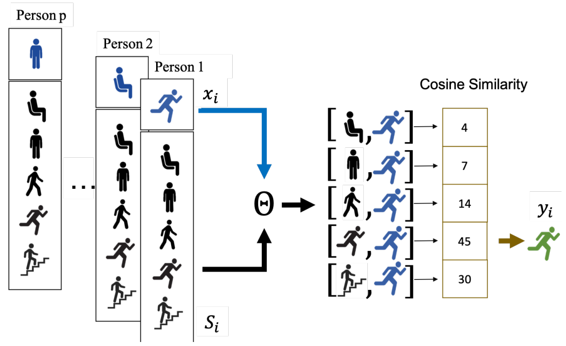


Fig. 3: Training MN^p for HAR; adapted from [13]

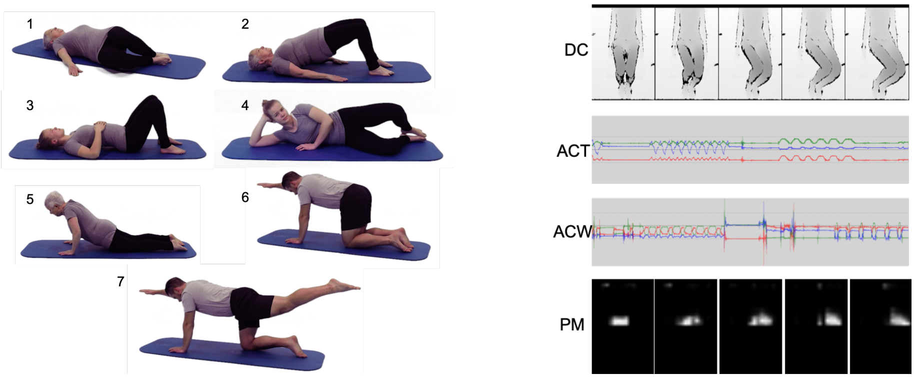
4 Evaluation and Results

Aim of the evaluation is two fold; firstly we analyse lower and upper bound performances of ExR algorithms using the 2 alternative methods of evaluation: person-agnostic versus person-aware, secondly we explore the transferability of personalised models for ExR from HAR. All experimental results calculate the mean F1-score and any significance is reported at the 95% level.

4.1 MEx Dataset

MEx is a sensor-rich dataset collected for 7 exercises with four sensors, publicly available at the UCI Machine Learning Repository[1]. Seven exercise classes are included this data collection; 1-Knee Rolling, 2-Bridging, 3-Pelvic Tilt, 4-Bilateral Clam, 5-Repeated Extension in Lying, 6-Prone Punch and 7-Superman (Figure4a). These exercises are frequently used for prevention or self-management of LBP.

[1] https://archive.ics.uci.edu/ml/datasets/MEx



(a) Exercises

(b) Raw sensor data

Fig. 4: MEx dataset

There are four sensor modalities; two accelerometers placed on the wrist and the thigh of the person; a pressure mat was where the person lays on to perform the exercises and a depth camera was placed above the person facing downwords. The tri-axial accelerometers record data at $100Hz$ frequency within the range of $\pm 8g$. The pressure mat and the depth camera record gray scale frames at $15Hz$ frequency and frame sizes are 32×16 and 240×320 respectively. Figure 4b shows a visualisation of each sensor data type.

In this study we focus on ExR with a single modality. Accordingly we create four datasets with the four modalities available on MEx; the thigh accelerometer, the wrist accelerometer, the pressure mat and the depth camera respectively referred to as ACT, ACW, PM and DC in the rest of this paper.

4.2 Pre-processing

The sliding window method is applied on each individual sensor data stream to create data instances; where the window size is 5 seconds with 3 seconds overlap. Each resulting window forms a data instance and is labelled with the exercise class. This process yields datasets of 6240 instances ($|\mathcal{X}| = 6240$), with 208 data instance per user in average ($|\mathcal{X}_p| \approx 208$). We apply a set of pre-processing steps for each sensor modality as recommended by the authors of [12]. A reduced frame rate of 1 frame/second is applied for DC and PM data and the DC data frames are compressed from 240×320 to 12×16. The inertial sensor data from ACW and ACT are pre-processed using the Discrete Cosine Transformation (DCT) according to [12].

4.3 Comparison of Person-agnostic and Person-aware Settings

In order to demonstrate the effect of different evaluation methods on ExR, we evaluate ExR algorithms using the two person-agnostic evaluation methodologies; R-HO and CF (Section 3.1) and the person-aware methodology LOPO.

We choose the best performing algorithms for each sensor dataset from [12] for our comparative study. We discard the user parameter on the data instances to obtain the person-agnostic datasets, and use 1/30 as the λ parameter to split a dataset for training and testing or to create folds. Accordingly we repeat the R-HO experiments for 30 iterations and perform 30 CF experiments. These are compared with results from person-aware LOPO from [12].

Methodology		Algorithm	ACT	ACW	DC	PM
Person-agnostic	R-HO	From [12]	0.9807	0.9163	0.9953	0.9905
	CF		0.9798	0.9260	0.9960	0.9880
Person-aware	LOPO	From [12]	0.9015	0.6335	0.8720	0.7408

Table 1: Mean F1-score results: person-agnostic vs. person-aware settings

In Table 1, there is a significant difference between the performance measures obtained with person-agnostic and person-aware methods. Inevitably when there is no person-wise disjoint train and test splits, algorithms have the opportunity to configure its parameters to better fit the expected user population at test time, resulting in significantly improved performance. It is noteworthy that both person-agnostic methods achieve similar mean F1-scores consistently with all four datasets. We highlight that person-aware LOPO performance measures set the “lower-bound” and person-agnostic performance measures set the “upper-bound” for the ExR task with each sensor modality.

4.4 Comparative Study of Non-personalised vs. Personalised Algorithms for ExR

Performance of non-personalised algorithms (from [12]) is compared with the personalised algorithm MN^p (from Section 3.3) for ExR. We evaluate with two person-aware evaluation methodologies; R-PHO and LOPO from Section 3.2. With R-PHO experiments a test set is formed with randomly selected instances from 1/3 of persons forming the set of test users (μ), and the train instances selected from the other 2/3 of train users. This is repeated for 10 test-train trials. In LOPO experiments, we select the set of data instances from one user as the test set and the rest forming the train set, and this is repeated 30 times but each time with a different test user. While R-PHO helps to evaluate transferability of the algorithms with multiple users at a time, LOPO evaluates the transfer to a single user at a time, both are valid scenarios for ExR and HAR in general. For comparative purposes, we also included results obtained by authors of [13] for general HAR tasks, pose detection tasks and Activities of Daily Living (ADL) classification task.

In Table 2 results are grouped under each task and the difference between the personalised algorithm and the best non-personalised algorithm is presented in the last column. Here the best non-personalised algorithm for the first three tasks is the non-personalised Matching Networks introduced in [11] and for the MEx exercises domain they are the best performing algorithms found by the authors of [12] for each sensor modality. Person-aware evaluation methodologies require a non-parametric statistical significance test as they produce results that are not normally distributed. We use the Wilcoxon signed-rank test for paired samples to evaluate the statistical significance at 95% confidence and highlight the significantly improved performances in bold.

Problem	Evaluation	Dataset	Algorithm		Difference
Domain	Methodology		non-personalised	MN^p	
Pose Detection	R-PHO	HDPoseDS_{17}	0.7678	**0.9837**	+21.68%
		HDPoseDS_{6}	0.4292	**0.9186**	+48.94%
General HAR	R-PHO	selfBACK$_{W,T}$	0.7340	**0.9169**	+18.29%
		selfBACK$_{W}$	0.6320	**0.8563**	+22.44%
ADL	R-PHO	PAMAP2	0.8715	0.8690	-0.25%
MEx	R-PHO	ACT	0.9064	**0.9424**	+3.60%
		ACW	0.6352	**0.6845**	+4.93%
		DC	0.8741	**0.9186**	+4.45%
		PM	0.6977	**0.8538**	+15.61%
MEx	LOPO	ACT	0.9015	**0.9155**	+1.40%
		ACW	0.6335	**0.6663**	+3.28%
		DC	0.8720	**0.9342**	+6.22%
		PM	0.7408	**0.8205**	+7.97%

Table 2: Mean F1-score Results for the comparison of non-personalised algorithm vs. personalised algorithm for ExR

The personalised algorithm, MN^p, has significantly outperformed the best non-personalised algorithm on both evaluation methodologies. Overall similar performance measures are observed across both methodologies, with the exception of the PM dataset, where there is a $\sim +5\%$ difference when using LOPO compared to R-PHO with the non-personalised algorithm.

We observe that personalisation has the greatest benefit for Pose detection, followed by general HAR tasks which feature both ambulatory and stationary activities; and thereafter on the ExR task. The exception here was ADL tasks, where personalisation neither improved nor degraded performance. Close examination of the activity duration of each domain suggests that pose and ambulatory activities are highly repetitive, or are performed in short repetitive time spans, which minimises the capturing of personal rhythmic nuances. In comparison, a single repetition of an exercise takes a longer time and consists of multiple sub steps making it harder to model due to potential variation opportunities between persons. Essentially MN^p is capable of finding some commonalities between exercise classes, but there is less opportunities than compared with pose or ambulatory movements. In contrast, ADL activities tend to have less clear start and stop demarcations and have even longer spans. They also have the highest possibility of featuring personal traits and nuances. For example, an ADL classes such as ironing or cleaning can be completely different from one user to another. Accordingly, personalised algorithms struggle to find such commonalities between ADL classes; explaining results we had observed here.

4.5 Distribution of Performance Measures

We visualise the distribution of performance measures with different methodologies for datasets ACT, ACW, DC and PM in Figures 5a to 5d. Each figure shows the distribution of results obtained from three experiments; CF method with the non-personalised algorithm using red triangles, LOPO method with the non-personalised algorithm using blue circles and LOPO method with the MN^p personalised algorithm using green stars.

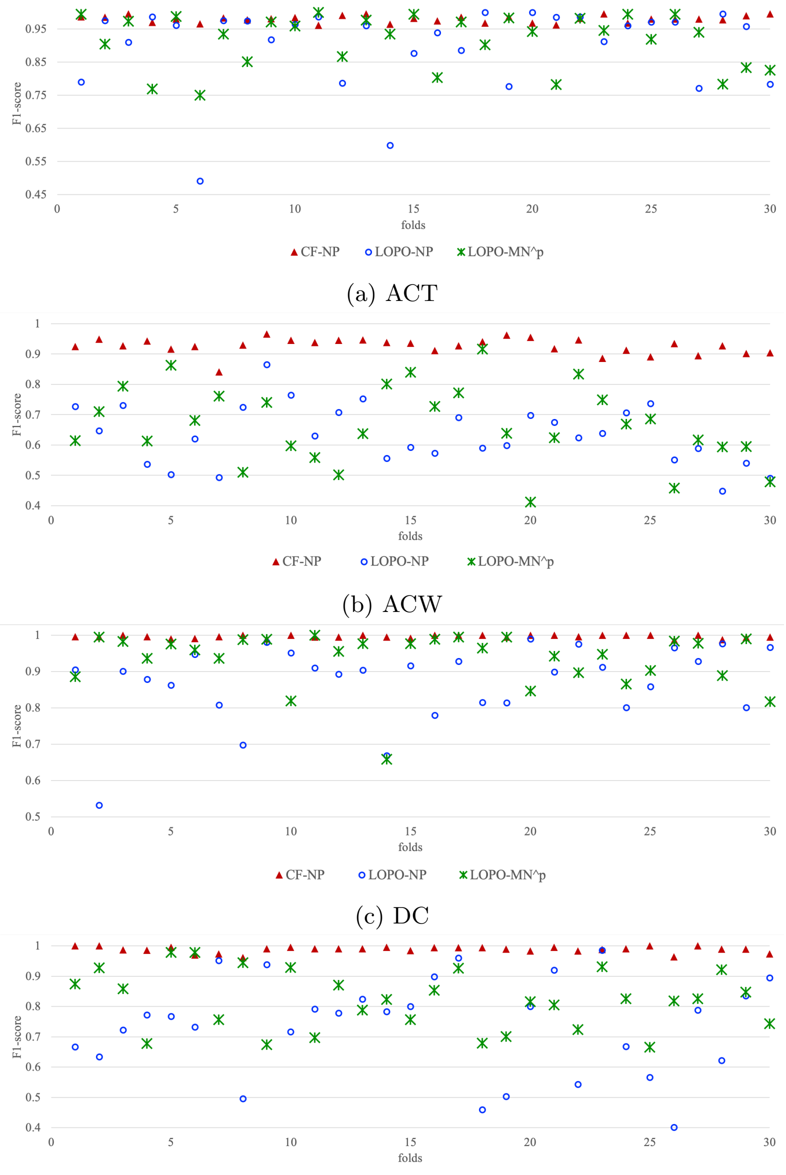
The CF results on every dataset highlight the upper-bound performance of the ExR task when evaluated under the assumption that the algorithm is trained and tested on the same user distribution. The LOPO results obtained for the non-personalised algorithm sets the lower-bound for the ExR task and highlights how non-personalised algorithms struggle when tested on a person that was not part of the training user group. Personalised algorithm such as MN^p, minimise the gap between the upper-bound and the lower-bound with a majority of users by improving upon the lower-bound. As shown in Table 2, MN^p, is a good choice for personalising across all tasks. With significant advantages shown with pose detection and fewest gains with the ExR task.

5 Conclusion

We present a comprehensive study of Exercise Recognition (ExR) model evaluation of adaptability to diverse user groups after deployment. We explore the different evaluation methods that share the same user set during training and testing (i.e. person-agnostic) and methods that keep disjoint user sets for training and testing (i.e. person-aware). We show how the prior method sets the upper-bound and latter sets the lower-bound for model performance in ExR. We highlight how person-agnostic evaluation results are normally distributed, but person-aware evaluation results are not, thus calling for non-parametric statistical significance testing methods. We adapted a personalised algorithm, MN^p, that is capable of learning a feature space that is transferable to unseen users and user groups and show how it outperforms the lower-bound while using the evaluation criteria suitable for model deployment(i.e. person-aware ExR). Finally we believe improving performance of ExR with personalised algorithms in the person-aware setting is a significant step towards deploying user-friendly, unobtrusive ExR algorithms with fitness applications. We identify the need to improve these personalised algorithms to better suit the ExR domain.

References

1. Berchtold, M., Budde, M., Gordon, D., Schmidtke, H.R., Beigl, M.: Actiserv: Activity recognition service for mobile phones. In: International Symposium on Wearable Computers (ISWC) 2010. pp. 1–8. IEEE (2010)
2. Burns, D.M., Leung, N., Hardisty, M., Whyne, C.M., Henry, P., McLachlin, S.: Shoulder physiotherapy exercise recognition: ML the inertial signals from a smartwatch. Physiological measurement **39**(7), 075007 (2018)
3. 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. IEEE (2010)
4. Miu, T., Missier, P., Plötz, 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. pp. 1138–1147. IEEE (2015)
5. Ohashi, H., Al-Naser, M., Ahmed, S., Nakamura, K., Sato, T., Dengel, A.: Attributes' importance for zero-shot pose-classification based on wearable sensors. Sensors **18**(8), 2485 (2018)
6. Qi, J., Yang, P., Hanneghan, M., Waraich, A., Tang, S.: A hybrid hierarchical framework for free weight exercise recognition and intensity measurement with accelerometer and ecg data fusion. In: 2018 40th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 3800–3804. IEEE (2018)
7. Sun, X., Kashima, H., Ueda, N.: Large-scale personalized human activity recognition using online multitask learning. IEEE Transactions on Knowledge and Data Engineering **25**(11), 2551–2563 (2013)
8. Sundholm, M., Cheng, J., Zhou, B., Sethi, A., Lukowicz, P.: Smart-mat: Recognizing and counting gym exercises with low-cost resistive pressure sensing matrix. In: Proceedings of the 2014 ACM UbiComp. pp. 373–382. ACM (2014)
9. Tapia, E.M., Intille, S.S., Haskell, W., Larson, K., Wright, J., King, A., Friedman, R.: Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In: 2007 11th IEEE international symposium on wearable computers. pp. 37–40. IEEE (2007)
10. Velloso, E., Bulling, A., Gellersen, H., Ugulino, W., Fuks, H.: Qualitative activity recognition of weight lifting exercises. In: Proc. 4th Augmented Human Int. Conf. pp. 116–123. ACM (2013)
11. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems. pp. 3630–3638 (2016)
12. Wijekoon, A., Wiratunga, N., Cooper, K.: Mex: Multi-modal exercises dataset for human activity recognition. arXiv preprint arXiv:1908.08992 (2019)
13. Wijekoon, A., Wiratunga, N., Sani, S., Cooper, K.: A knowledge-light approach to personalised and open-ended human activity recognition. Knowledge-Based Systems p. 105651 (2020)
14. Xiao, F., Chen, J., Xie, X.H., Gui, L., Sun, J.L., none Ruchuan, W.: Seare: A system for exercise activity recognition and quality evaluation based on green sensing. IEEE Transactions on Emerging Topics in Computing (2018)
15. Zhou, B., Sundholm, M., Cheng, J., Cruz, H., Lukowicz, P.: Never skip leg day: A novel wearable approach to monitoring gym leg exercises. In: IEEE Int. Conf. Pervasive Computing and Communications. pp. 1–9. IEEE (2016)



(a) ACT

(b) ACW

(c) DC

(d) PM

Fig. 5: Distribution of F1-score over the folds