Journal of Reliable Intelligent Environments (2023) 9:27–39
https://doi.org/10.1007/s40860-021-00167-w

ORIGINAL ARTICLE

Deep learning and model personalization in sensor-based human activity recognition

Anna Ferrari[1] · Daniela Micucci[2] · Marco Mobilio[2] · Paolo Napoletano[2]

Received: 11 June 2021 / Accepted: 22 December 2021 / Published online: 7 January 2022

Abstract

Human activity recognition (HAR) is a line of research whose goal is to design and develop automatic techniques for recognizing activities of daily living (ADLs) using signals from sensors. HAR is an active research filed in response to the ever-increasing need to collect information remotely related to ADLs for diagnostic and therapeutic purposes. Traditionally, HAR used environmental or wearable sensors to acquire signals and relied on traditional machine-learning techniques to classify ADLs. In recent years, HAR is moving towards the use of both wearable devices (such as smartphones or fitness trackers, since they are daily used by people and they include reliable inertial sensors), and deep learning techniques (given the encouraging results obtained in the area of computer vision). One of the major challenges related to HAR is population diversity, which makes difficult traditional machine-learning algorithms to generalize. Recently, researchers successfully attempted to address the problem by proposing techniques based on personalization combined with traditional machine learning. To date, no effort has been directed at investigating the benefits that personalization can bring in deep learning techniques in the HAR domain. The goal of our research is to verify if personalization applied to both traditional and deep learning techniques can lead to better performance than classical approaches (i.e., without personalization). The experiments were conducted on three datasets that are extensively used in the literature and that contain metadata related to the subjects. AdaBoost is the technique chosen for traditional machine learning, while convolutional neural network is the one chosen for deep learning. These techniques have shown to offer good performance. Personalization considers both the physical characteristics of the subjects and the inertial signals generated by the subjects. Results suggest that personalization is most effective when applied to traditional machine-learning techniques rather than to deep learning ones. Moreover, results show that deep learning without personalization performs better than any other methods experimented in the paper in those cases where the number of training samples is high and samples are heterogeneous (i.e., they represent a wider spectrum of the population). This suggests that traditional deep learning can be more effective, provided you have a large and heterogeneous dataset, intrinsically modeling the population diversity in the training process.

Keywords Human activity recognition · Personalization · ADL · Machine learning · Deep learning · AdaBoost · Convolutional neural network

1 Introduction

According to the World Report on Ageing and Health [25], expectancy of life has dramatically increased in the last decades. However, these additional years are often characterized by the presence of diseases, whose symptoms must be constantly monitored to prevent a worsening of the clinical situation (e.g., neurodegenerative and non-communicable diseases). As a result, the aging population causes an increase in the number of hospital admissions and resources needed for care, including rehabilitation. Despite that, a pure human-based monitoring becomes unsustainable as the rate between

✉ Daniela Micucci
daniela.micucci@unimib.it

Anna Ferrari
a.ferrari@cern.ch

Marco Mobilio
marco.mobilio@unimib.it

Paolo Napoletano
paolo.napoletano@unimib.it

[1] CERN, 1211 Geneva 23, Switzerland

[2] University of Milano-Bicocca, Viale Sarca 336, Milan, Italy

the elderly and working population is decreasing dramatically.

An automatic and remote monitoring system of human behavior can help with the problem of population aging: on one hand, it substantially reduces the healthcare costs, and on the other hand, it improves the patients' life quality and their independence. However, transforming a human-based monitoring system to an automatic-based human activity recognition system is not a simple task. Recognizing an activity or understanding a situation are relative easy for humans but become extremely complex for computers as they require sophisticated techniques for data preprocessing and analysis.

The automatic recognition of the human activities is known as human activity recognition (HAR), a relatively young research area that is attracting more and more researchers thanks to the significant technological advances.

Several methods have been defined over the years for HAR to improve care capability and efficiency and demonstrate high potential to improve diseases prevention, remote monitoring, and smart diagnosis for elderly.

Environmental and wearable devices play a major role in such methods [19]. However, given the inherent limitations of environmental sensors (e.g., intrusiveness, costs, limitation to instrumented environments), research is shifting towards the use of wearable devices equipped with inertial sensors. In recent years, smartphones have gained increasing interest: they are equipped with several sensors able to capture attributes of interest as motion, location, temperature, and ECG; they are part of the daily life of the people and thus do not require any change of people's behavior; they have a wide worldwide spread; and finally they can be used both indoor and outdoor.

In addition to the advantages highlighted above, what makes the use of smartphones attractive is their increasing power that makes them almost comparable to laptops: nowadays smartphones acquire, store, share, and elaborate huge amounts of data in a very short time while also preserving energy power.

Initially, traditional machine-learning techniques were used for sensor-based HAR [4]. The most used techniques were discriminative analysis (DA), Naive Bayes (NB), support vector machine (SVM), hidden Markov models (HMM), joint boosting (JB), AdaBoost, and k-NN [4,19]. Traditional machine learning methods (ML) are low cost in terms of time consumption, data, and complexity. However, their dependency on expert knowledge in the features extraction phase often leads to the generation of models that are expensive (require an expert) and difficult to compare [12,41,43]. On the other side, deep learning methods (DL) remain stable in terms of feature extraction, which is mainly automatically executed, but the training phase requires more data, and, consequently, it is either time consuming or requires expansive hardware [7].

Regardless of the underlying learning method (either traditional machine learning or deep learning), HAR techniques do not achieve satisfying recognition accuracy in real-world applications. Indeed, HAR techniques struggle to generalize to new users and/or new environments [16,17], mainly because of the population diversity problem [18]: people perform the same activities differently. According to Zunino et al. [44], two factors are the main cause of the population diversity problem: the anthropometric differences of body parts or the incongruous personal styles in accomplishing the scheduled action (termed *inter-subject variability*); and the diversity with which a subject carries out the same action (termed *intra-subject variability*).

To face subjects variability, algorithms should be trained on a representative number of subjects and on as many cases as possible. The number of subjects in the dataset does not just impact the quality and robustness of the induced model, but also the ability to evaluate the consistency of results across subjects [21].

Another way to face variability is to consider *similarity* between subjects and signals as a key factor to obtain more robust recognition models. Subjects with similar physical characteristics perform activities relying on similar patterns (*physical-based similarity*); subjects, even are physically dissimilar, can perform activities relying on similar patterns (*signal-based* or *sensor-based similarity*). Personalization applied to traditional machine-learning techniques results in robust activity recognition models [10].

The positive results obtained by applying personalization to traditional machine-learning techniques and the well-known ability of deep learning techniques to generalize, led us to experiment personalization also on deep learning techniques. Our research aimed to answer the following research questions.

- *Does personalized deep learning outperform personalized machine learning?*
- *Does deep learning outperform personalized machine and deep learning?*

We have started to investigate the benefits of personalization applied to traditional machine-learning techniques [10]. The results obtained will be used in this paper to compare the techniques with each other (traditional machine learning and deep learning with personalization and without personalization). From what concerns personalized deep learning techniques, preliminary results were presented at a workshop on Artificial Intelligence for an Ageing Society [11]. Since the results did not allow us to come to a confident conclusion, we extended the experimentation.

The evaluation has been performed on three public domain datasets (i.e., UniMiB SHAR [24], Motion Sense [22], and MobiAct [37]), because they were acquired from smart-

phones and to the best of our knowledge, they are among the few that include additional information about the subjects' characteristics [8]. We use AdaBoost as a traditional machine-learning classifier because it permits to weight input data according to subject and sensor similarities, and also because it is one of the most performing classifiers [26,38]. Finally, we use a Residual Network (ResNet, a CNN-based technique) as a deep learning technique, which is based on the traditional architecture proposed by He et al. [15] which demonstrated to be very effective.

The obtained results show that personalization applied to ResNets leads to more accurate models with respect to the ones obtained by applying personalization to AdaBoost only in one dataset, namely Motion Sense. Traditional ResNets in average obtained better results in most of the configurations used. Moreover, a regular ResNet without personalization performs, in most of the cases, better than a personalized ResNet, thus demonstrating that population diversity can be taken into account using a large variety of data and a robust deep learning technique.

We can summarize the main contributions of our research as follows.

- Definition of a personalization strategy to be applied to both traditional machine learning and deep learning. Personalization allows to build recognition models that may take advantage of additional information, that is, the physical characteristics of the subjects and the signals.
- Empirical evaluation of the performance obtained by applying personalization to both machine-learning and deep learning techniques.
- Empirical comparison of the performance obtained from traditional machine-learning and deep learning models with both personalization and without personalization.

The paper is organized as follows. Section 2 discusses relevant literature related to personalization in HAR; Sect. 3 discusses similarity and specifies how it is employed in traditional machine- and deep learning techniques; Sect. 4 describes the setup of our experiments; Sect. 5 presents the results of the experiments; finally, Sect. 6 presents the conclusions and outlines future research on personalization.

2 Related work

Automatic human activity recognition is a complex task. Algorithms struggle to generalize to new users and environment, thus requiring a significant effort when they have to cope with the real world [16,17].

One of the main challenges is related to the population diversity problem [18], that is, the natural differences between users' activity patterns, which implies that different executions of the same activity are different.

To try to solve the population diversity problem, datasets used to train the models should contain a representative number of subjects (to face with the inter-subject variability) and a significant number of signals from the same subject (to face with the intra-subject variability) [21]. Unfortunately, generating datasets with these characteristics is not a simple task. This difficulty is also reflected in the datasets that are currently available (acquired via wearable devices) as reported in recent surveys [8,32].

To achieve the desired level of generality of the trained models basing on the actual available datasets, researchers have recently started experimenting with personalization-based techniques.

Personalization is approached differently in the literature depending on whether it is applied to machine-learning or deep learning techniques.

What the techniques share and that does not depend neither on the personalization strategy adopted nor on the type of technique (machine vs. deep learning) are the methods with which the dataset is divided into train and test sets. There are mainly three approaches to split the samples into the two sets (train and test) [10]. They differ in the way they use the samples from the end-user in the train set: the subject-independent approach does not include the samples of the end-user in the train set; the subject-dependent approach includes the samples of the end-user both the train and the test sets; finally, the hybrid approach includes the samples of both the end-user and of other users in both the train and the test sets. The first two approaches where initially identified by Tapia et al. [35]. Some years later, Weiss at al. [40] also introduced the hybrid approach.

The approaches were compared to identify the one that yields better performance. For examples, Medrano et al. [23] and Shen et al. [29] compared the subject-dependent and subject-independent approaches and concluded that the subject-dependent approach performs better. Other researchers who compared the three methods came to the conclusion that models that rely on the subject-dependent and the hybrid approaches outperform the performance of the models based on the subject-independent approach [6,20,36,40].

Personalized machine-learning and deep learning techniques differ mostly in the way the end-user enters the model generation process. Personalized machine-learning techniques manly use user info (such as physical characteristics) and the context (such as the device position) to generate the model. Personalized deep learning techniques rely on the availability of additional samples of the end-user to update and slightly modify pre-trained models.

The following two sub-sections describe the proposed approaches for machine-learning and deep learning techniques, respectively, that rely on personalization.

2.1 Personalized machine learning

Personalized machine-learning techniques can be divided o into two major categories: similarity-based approaches and classifier-based approaches.

Similarity-based approaches personalize the recognition model by exploiting the similarity between signals and/or between end-user characteristics. For example, Sztyler et al. [33,34] consider the similarity between signals and propose to train the models only with subjects with signals similar to those of the end-user. Ferrari et al. [10] experiment with the joint use of similarity between signals and between physical characteristics of end-users. Lane et al. [18], on the other hand, propose an approach that weights samples differently based on the level of similarity of both signal and physical characteristics to the end-user. Finally Garcia-Ceja et al. [13,14] exploit inter-class similarity, thus training the model using only the instances that are similar to the end-user for each class.

Classifier-based approaches personalize the recognition model by combining several activity recognition models. For example, Hong at al. [16] propose to combine models that have been trained relying on an a subject-dependent approach for dataset split. Reiss et al. [27] propose a model that relies on a set of weighted classifiers.

2.2 Personalized deep learning

There are two main techniques used in the deep learning context to personalize models: transfer learning and incremental learning.

Transfer learning approaches update a pre-trained network by calibrating the weights when a new user enters the stage. This approach is particularly useful when labeled data of the end-user are not available at the time the model is generated. Rokni et al. [28] exploit transfer learning by updating a CNN model trained with data collected from a few participants. The model is updated by fine-tuning the top layers of the CNN with a small amount of data of the end-user.

Incremental learning approaches are based on the update of a pre-existing model when new data become available, including data from previously unseen users [31]. For example, Yu et al. [42] obtain a personalized model by first training the model using a subject-independent approach and then incrementally updating it with the samples of the end-user by assigning them a major weight. Siirtola et al. [30] propose an incremental learning method combined with Learn++, which has been augmented by Amrani et al. [2] with deep learning. A similar approach is proposed by Vo et al. [39]. Finally, Abdallah et al. [1] exploit clustering to tailor the model to a specific user.

3 Proposed methods

This section introduces the concept of similarity and how it is applied to personalize traditional machine-learning and deep learning models.

3.1 Similarity

To take into account the population diversity, we introduce the concept of similarity between subjects. Similarity may be used to properly weight the training data to give more importance to data that are more similar to data of the user under test. Similarity derives from two basic intuitions [10, 18].

1. Two individuals who have similar physical characteristics are expected to generate similar signals from inertial sensors when performing the same activities.
2. Although users have different physical characteristics, they may generate similar signals from inertial sensors when performing the same activities.

To evaluate similarity between subjects, we describe each subject i with a feature vector:

$$\mathbf{g}_i = g_1, \ldots, g_K. \tag{1}$$

Then, similarity between the subject i and another subject j is defined as follows:

$$sim(i, j) = e^{-\gamma d(i,j)}, \tag{2}$$

where γ is a scale parameter and $d(i, j)$ is the Euclidean distance between the feature vectors of the subjects i and j:

$$d(i, j) = \sqrt{\sum_{k=1}^{K} (g_{k,i} - g_{k,j})^2}. \tag{3}$$

The resulting similarity value ranges from 0 to 1, where 0 means that the subject i is dissimilar to subject j, and 1 means that subjects i and j are similar.

The two basic intuitions lead to identify two types of similarity between subjects.

– Physical-based similarity, that is, the similarity based on the physical characteristics of the subjects and which derives from the first intuition.

Table 1 Hand-crafted features used in signal-based similarity

Feature	Definition
Minimum	$\min = \min_{j=1,\ldots n}(x_j)$
Maximum	$\max = \max_{j=1,\ldots n}(x_j)$
Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Median	$\text{Me} = x_{0.5} : F(x_{0.5}) = 0.5$
Standard deviation (SD)	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
Variance	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Interquartile difference	$\text{ID} = x_{0.75} - x_{0.25}$
Skewness	$S = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$
Kurtosis	$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4}$
Root mean square (RMS)	$\text{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$
Total sum	$\text{TS} = \sum_{i=1}^{n} x_i$
Range	$R = \max - \min$
Mean of peak's distance	$m_p = \frac{1}{s^2}\sum_{j=1}^{s}\sum_{i=1}^{s} d(p_i, p_j)$
Entropy	$H(x) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$
Sum of the spectral power components	$\text{SP} = \frac{1}{n}\sum_{j=1}^{f} \lvert\text{FFT}_j\rvert^2$
Mean of the spectral components	$\mu_f = \frac{1}{n}\sum_{j=1}^{n} \text{FFT}_j$
Median of the spectral components	$\text{Me}_f = \text{FFT}_{0.5} : F(f_{0.5}) = 0.5$
First cepstral coefficient	$c(1) = \mathcal{F}^{-1}\{\log \lvert A(f)\rvert\}$, where $A(f)$ is the Fourier transform

– Signal-based (or sensor-based) similarity, that is, the similarity between the signals originated by two subjects which derives from the second intuition.

In the case of physical-based similarity ($\text{sim}^{\text{physical}}$), the feature vector in Eq. 1 will consist of values that describe the physical characteristics of the individual. The selection of the physical characteristics has been inspired by the related literature and it is constrained to the availability of the metadata in the public datasets. Thus, we define a feature vector as composed by three real values $\mathbf{g}^{\text{physical}} = (\text{age}, \text{weight}, \text{height}) = (g_1^p, g_2^p, g_3^p)$. Each component of the triplet ranges from 0 to 1 because all the ages, weights, and heights of the subjects have been normalized to fit the range of real number $[0 - 1]$.

In the case of signal-based similarity ($\text{sim}^{\text{sensor}}$), the feature vector in Eq. 1 will consist of the 18 features listed in Table 1. For each subject, we define a feature vector that is made of 18 real values described in Table 1: $\mathbf{g}^{\text{sensor}} = (g_1^s, \ldots, g_{18}^s)$. Each subject i has N_i segments. We calculate the similarity between 2 subjects i and j by summing the similarity between each segment of the subjects:

$$\begin{aligned}\text{sim}^{\text{sensor}}(i, j) &= \sum_{m=1}^{N_i}\sum_{n=1}^{N_j}\text{sim}(x_{in}, x_{jm}) \\ &= \sum_{m=1}^{N_i}\sum_{n=1}^{N_j} \text{e}^{-\gamma d(x_{in}, x_{jm})}. \end{aligned} \tag{4}$$

Finally, we identify a third kind of similarity that combines physical with sensor-based similarity ($\text{sim}^{\text{physical+sensor}}$). This similarity is obtained as the weighted sum of physical and sensor similarity:

$$\begin{aligned}&\text{sim}^{\text{physical+sensor}}(i, j) \\ &\quad = \alpha \cdot \text{sim}^{\text{sensor}}(i, j) + \beta \cdot \text{sim}^{\text{physical}}(i, j),\end{aligned} \tag{5}$$

where α and β are such that $\alpha + \beta = 1$.

3.2 Personalizing the methods

As previously introduced, the idea is to take advantage of the similarity between subjects in machine-learning and deep learning engines: the similarity between subjects is, respectively, used to weight and to select the training data to give higher priority to data belonging to the most similar subjects to the target one.

In particular, for what concerns personalization in traditional machine learning (PML), we consider the similarity between users by augmenting the training data with the weights obtained by computing the similarity as defined in the previous section. Thus, the classification is influenced by the similarity between users.

Personalization in deep learning (PDL) considers a different approach. We identify a minimum value $\hat{m} > 0$ of subjects m and we use this value to select the most $\hat{m}$ similar subjects with respect to the target subject. Thus, the network is trained with the samples belonging to these $\hat{m}$ subjects. To evaluate the effect of the choice of m on the goodness of the network, we explored all the m values from the minimum value $\hat{m}$ to the maximum value which corresponds to the maximum number of subjects available in the dataset.

4 Experimental setup

This section illustrates the configuration and data we adopted in our experiment phase. The ultimate goal is to evaluate if personalization applied to deep learning techniques (PDL) allows to obtain more robust HAR models with respect to the ones obtained by both personalized machine-learning techniques (PML) and deep learning techniques (DL).

Table 2 Convolutional neural networks hyperparameters' settings

Layer name	Shape
Convolutional	{1×3}
Activation	ReLu
Max pooling	{1×3}
Dropout	0.9
Fully connected	148× f_{maps}
Softmax	1× num classes

4.1 Implementation details

To build a PML model, we considered the AdaBoost classifier which permits to weight training data before starting the training process [3,10].

To build PDL and DL models, we rely on Convolutional Neural Networks. In particular, we selected a Residual Network (ResNet) based on the ResNet proposed in [7] and [9].

Table 2 details the network used in our study. The input size of the network is $1 \times 128 \times 3$, that corresponds to three segments along the three axes x, y, and z. The network architecture is made of an initial convolutional block, 3 residual stages (each containing a variable number n of residual blocks), an average pooling layer, a fully connected layer, and a softmax layer. A convolutional block is made of three layers: convolutional, batch normalization, and ReLu. A residual block is made of 2 subsequent convolutional blocks and an additional operator that sums the input of the residual block with the output of the residual block itself. Each convolutional layer is $1 \times 3 \times f_{maps}$, where f_{maps} is the number of feature maps of the filter. For each dataset, the best values for n and f_{maps} have been found by following a grid search approach: n ranged between 3 and 21, while f_{maps} ranged between 10 and 200.

For what concerns the personalization in PDL, we selected the parameter $\hat{m}$ equals to 10. The network has been initially trained using data from the selected $\hat{m} = 10$ subjects and then we added five subjects until the maximum number of subjects in the dataset is achieved.

4.2 Datasets

We considered three public datasets containing accelerometer signals of activities of daily living (ADLs) recorded with smartphones. The selected datasets are the same as those discussed in [10]: the article describes activities, sample distributions, and other useful information related to the labeled samples. Each dataset includes gender, age, weight, and height of each subject.

Table 3 Number of segments divided into training, validation, and test dataset and the number of the classes for each dataset

Dataset	#train	# validation	# test	# classes
MobiAct	34,070	9734	4867	15
Motion Sense	14,945	4270	2135	6
UniMiB SHAR	8240	2354	1177	17

- UniMiB SHAR [24] contains tri-axial acceleration data organized in 3 s windows around the peak. The dataset contains 17 different activities (both ADLs and Falls) performed by 30 subjects. The sampling rate is 50 Hz. We have chosen segments of 3 s for this dataset. The subjects placed the smartphone used for the acquisition (a Samsung Galaxy Nexus I9250) half of the times in the left trouser pocket and the remaining times in the right one.
- Motion Sense [22] contains time-series data generated by the accelerometers in an iPhone 6s worn by 24 participants. Each of the subjects performed 6 activities (only ADLs). The smartphone were kept in the participant's front pocket. The sampling rate is 50 Hz. We have chosen segments of 5 s for this dataset.
- MobiAct [37] includes tri-axial acceleration data of 15 ADLs and Falls recorded with a Samsung Galaxy S3 and performed by 67 participants. The windows size we considered is of 5 s with a sample rate of 87 Hz. The smartphone is located with random orientation in a loose pocket chosen by the subject.

4.3 Data split

Data have been split according to two different configurations [10]: subject-independent (SI) and hybrid (HYB). The SI data split configuration does not use the end user data for the development of the activity recognition model, that is, the classification model is trained on the data of the users except the end-user. The HYB data split configuration uses the end-user data and the data of the other users for the development of the activity recognition model, that is, the classification model is trained both on the data of the users and on a part of the data of the end-user.

We do not employ the subject-dependent split because by definition, it does not contain samples from other users and then it is not possible to compute the similarity between different subjects.

4.4 Metrics

We measure the algorithms performance in terms of average accuracy, that is, given E the set of all the activities types,

$a \in E$, NP_a the number of times a occurs in the dataset, and TP_a the number of times the activity a is recognized; accuracy is defined as in the following equation:

$$\text{Acc} = \frac{1}{|E|}\sum_{a=1}^{|E|}\text{Acc}_a = \frac{1}{|E|}\sum_{a=1}^{|E|}\frac{\text{TP}_a}{\text{NP}_a}. \quad (6)$$

Acc is the arithmetic average of the accuracy Acc_a of each activity.

5 Results and discussion

Our study aims at providing answers to the following two research questions.

- RQ1: does personalized deep learning outperform personalized machine learning? In a previous research, we proved that personalized machine-learning methods outperform traditional machine-learning methods using a hybrid split [10]. With this research question, we investigate whether the application of personalization to deep learning techniques allows to achieve better accuracy with respect to personalized machine-learning techniques.
- RQ2: Does deep learning outperform personalized machine and deep learning? Deep learning techniques are proving to be effective also applied to the recognition of ADLs from inertial signals for many reasons as discussed in [5]. This research question investigates whether personalization applied to traditional and deep learning techniques can lead to better results than those obtained relying only on deep learning techniques.

5.1 RQ1: does personalized deep learning outperform personalized machine learning?

Table 4 shows the accuracy achieved by personalized deep learning (PDL) and personalized machine-learning (PML) methods for each dataset, data split, and type of similarity.

Accuracies are grouped by dataset (column Dataset) and then by configuration (column Model): split type (subject-independent and hybrid; in Table 4 referred as SI and Hyb, respectively) and then similarity type (physical, sensor, and physical in combination with sensor; in Table 4 referred as phy, sen, and phy+sen, respectively).

Columns three to seven (columns PDL) show the accuracy with respect to the values of m by applying personalization to deep learning (PDL). As introduced in Sect. 4, the number m represents the number of the most similar subjects compared to the test subject in terms of physical, sensor, and physical combined with sensor attributes. We recall that our experimentation starts with $m = 10$ and then it is increased by 5 elements at a time until all available subjects have been included. Due to the high number of subjects in the MobiAct dataset, the results referred to $m = 30, 35, 40, 45, 50, 55$ have been grouped together and the minimum and maximum accuracy are shown (column PDL >= 30).

Last column (column PML) shows the accuracy obtained by applying the personalization to machine learning (PML). Accuracies have been calculated as the average over the subjects.

First, we discuss PDL performances between the datasets and afterwards we compare PDL with PML.

5.1.1 Analysis of the performance of PDL between the datasets

To analyze the performance between datasets, we chose the hybrid data split configuration because, as shown in Table 4, it enables better results than the subject-independent one.

Table 4 also shows that the MobiAct dataset achieves better performance using PDL models in comparison with the other two datasets. This result is due to the size of the training dataset: when all subjects are considered, the MobiAct dataset has the largest training dataset size that includes 12,400 samples for each split. In contrast, UniMiB SHAR has up to 6800 training samples and Motion Sense up to 8000.

Figure 1 relates the size of the datasets to the accuracy of the PDL models. In particular, Fig. 1 shows the accuracy obtained by PDL models on the three datasets in the case of the hybrid data split and with the three types of personalization (physical, sensor, and the physical and sensor combination). The x-axes show the value of the m parameter. The left y-axes show the accuracy $\pm$ standard deviation of the hybrid model, whereas the right y-axes show the number of samples.

The orange barplot represents the frequency distribution of the total number of the samples belonging to the training dataset, with respect to the number of subjects m. The blue line shows the accuracy of the models with respect to the value of the m parameter.

If we compare the PDL performance for the same number of subject m, we observe that MobiAct has in general less training samples in comparison with UniMiB SHAR and Motion Sense. In particular, we have the following results varying the value of the m parameter.

- $\hat{m} = 10$, MobiAct has 2146 training samples with an accuracy of 85.23%, UniMiB SHAR has 2705 training samples with an accuracy of 46.17%, and Motion Sense has 3472 with an accuracy of 78.79%.
- $\hat{m} = 15$, MobiAct has 3263 training samples with an accuracy of 86.32%, UniMiB SHAR has 3837 training

Table 4 Experimental results—accuracy of personalized deep learning (PDL), personalized machine learning (PML)

Dataset	Model	PDL					PML
*m*th nearest subjects		10	15	20	25	>=30	57
						Min–max	
MobiAct	SI-phy	75.88	78.96	81.54	82.59	83.02–**86.08**	81.62
	SI-sen	71.75	74.36	76.25	77.68	77.42–80.14	**83.45**
	SI-phy+sen	71.88	74.11	75.97	77.38	78.45–79.68	**82.64**
	Hyb-phy	75.75	76.51	77.67	78.77	79.43–81.04	**89.43**
	Hyb-sen	78.15	78.77	79.45	80.39	80.90–81.40	**90.76**
	Hyb-phy+sen	85.23	86.32	86.96	87.40	87.58–88.17	**90.90**
Average						82.75	**86.46**
*m*th nearest subjects		10	15	20	25	–	27
UniMiB SHAR	SI-phy	25.49	27.61	31.48	35.42		**57.39**
	SI-sen	40.71	42.14	42.65	42.83		**57.00**
	SI-phy+sen	41.02	42.21	42.50	42.66		**56.93**
	Hyb-phy	42.87	43.69	45.33	45.82		**85.44**
	Hyb-sen	47.26	45.99	46.77	46.49		**84.71**
	Hyb-phy+sen	46.17	46.77	46.77	45.39		**84.87**
Average					43.46		**71.05**
*m*th nearest subjects		10	15	20	–	–	22
Motion Sense	SI-phy	74.30	77.40	**78.02**			72.45
	SI-sen	75.91	77.83	**78.80**			74.03
	SI-phy+sen	75.77	77.76	**79.00**			73.85
	Hyb-phy	77.59	79.44	**80.17**			77.76
	Hyb-sen	78.51	80.08	**80.38**			78.06
	Hyb-phy+sen	78.79	80.25	**80.41**			77.86
Average				**79.46**			75.66

SI subject-independent, *Hyb* hybrid, *phy* physical, *sen* sensor, *phy+sen* physical and sensor
For each row, the values in bold correspond to the best accuracy obtained with respect to a dataset and a configuration (column Model), when varying the technique (PDL and PML) and the configurations of the *m* parameter

samples with an accuracy of 46.77%, and Motion Sense has 5280 with an accuracy of 80.25%.
– $\hat{m} = 20$, MobiAct has 4345 training samples with an accuracy of 86.96%, UniMiB SHAR has 5020 training samples with an accuracy of 46.77%, and Motion Sense has 6952 with an accuracy of 80.41%.

In the case of UniMiB SHAR, differences in training data size have not a relevant influence in the models performance. From $\hat{m} = 10$ to $\hat{m} = 20$, the accuracy is not improved. That is because of the highest similarity between the subjects in the dataset.

Nevertheless, given m, PDL models perform much better on Motion Sense and on MobiAct in comparison with UniMiB SHAR.

In the case of MobiAct and Motion Sense datasets, the training size seems to do not have the same influence. Indeed, even if MobiAct presents less training samples, it outperforms Motion Sense. This behavior is certainty due to the subject's similarity in the specific dataset.

Similarities play a relevant role in PDL models performances. In particular, MobiAct presents a more variable similarity matrix compared to UniMiB SHAR and Motion Sense. Figure 2 shows the similarity matrices of the three datasets according to the three similarities: physical, sensor, and the combination between physical and sensor.

It is possible to notice that for all kind of similarity, UniMiB SHAR presents very low differences. The parameter γ is equal to 1. This choice allows us to compare the effective similarity between subjects. UniMiB SHAR shows high similarity over subjects, in comparison with MobiAct and Motion Sense where there exists more variability. It results that the more the differences between users, the more the generalization capability of the algorithm.

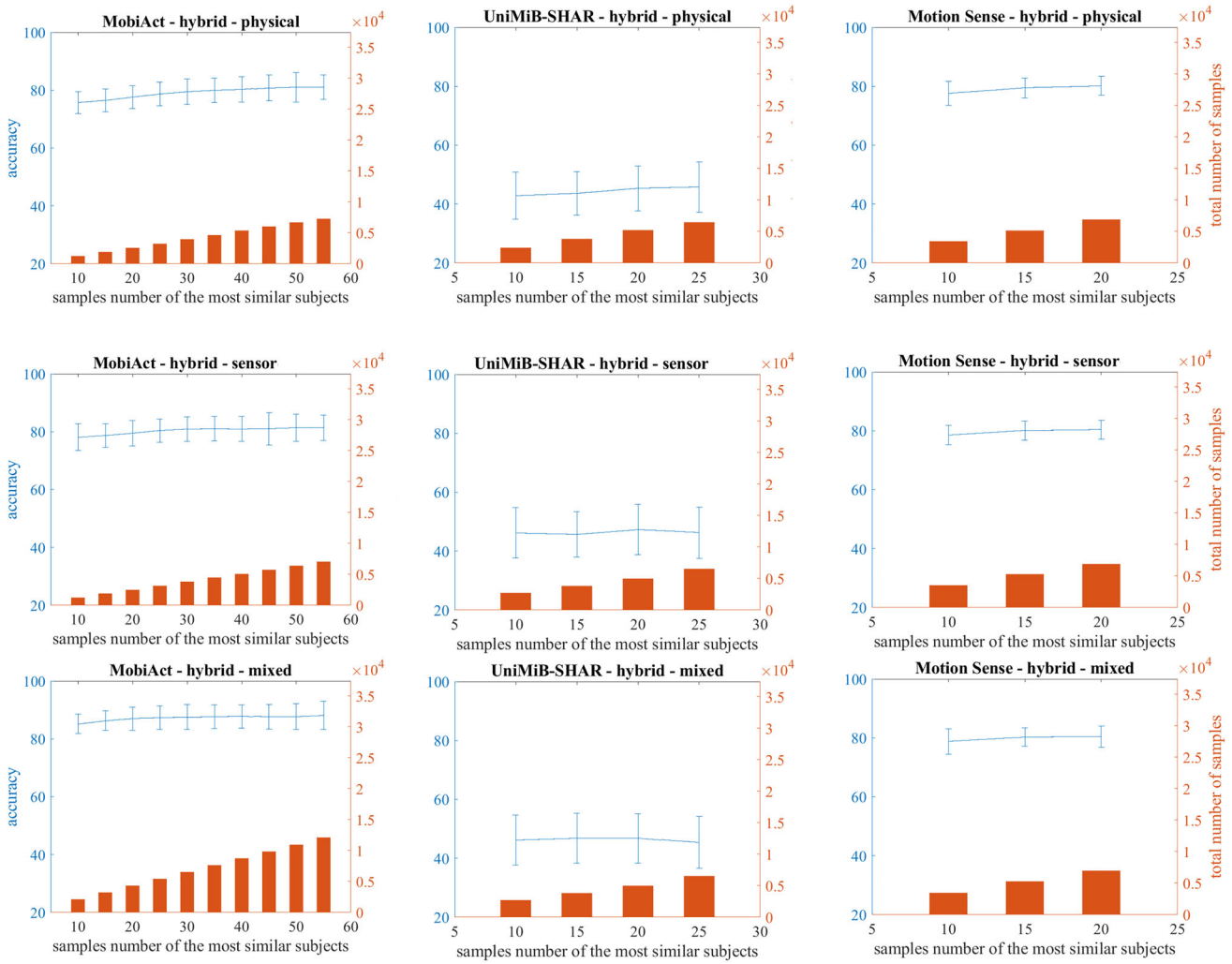

Fig. 1 Hybrid PDL models performances with different values for the m parameter (blue line) ± standard deviation and sample frequency distribution (orange bars)

5.1.2 Analysis of performance of PDL with respect to PML

Table 4 highlights that in the case of the MobiAct and UniMiB SHAR datasets, PML models overcome PDL ones, except for one case only. Namely, when it is used a personalization based on physical characteristics, a subject-independent data split, and the parameter m is greater than 20.

In the case of MobiAct dataset, the best performance is achieved using hybrid data split with the combination of physical and sensor attributes (accuracy equal to 90.90%). In average, PML achieves 86.46% of accuracy, about 4% more than PDL accuracy.

In the case of the UniMiB SHAR dataset, PML models achieve better performance than PDL in all of the cases. The best accuracy equals to 84.47% and also corresponds to hybrid data split with the combination of physical and sensor attributes. In this case, the margin with respect to the corresponding PDL is of 38.10%. In average, PML accuracy achieves 71.05%, while PDL only 43.46%, with a margin of 27.59%.

Motion Sense dataset shows a completely different behavior. The accuracy of the PDL models always outperform the PML accuracy. The best model corresponds to the hybrid data split with the combination of physical and sensor attributes, which reaches the 80.41% of accuracy. The corresponding PML models achieve 77.86% by a margin of 2.55%. In average, PDL models achieve an accuracy of 79.46% by a margin of 3.8% to PML.

For MobiAct and Motion Sense datasets, the difference between PDL and PML models is not relevant. On the opposite, in the case of UniMiB SHAR, PML models provide a relevant contribution to the classification performance, by a margin of about 27%. This difference is explained, once again, by looking at the similarity matrix in Fig. 2.

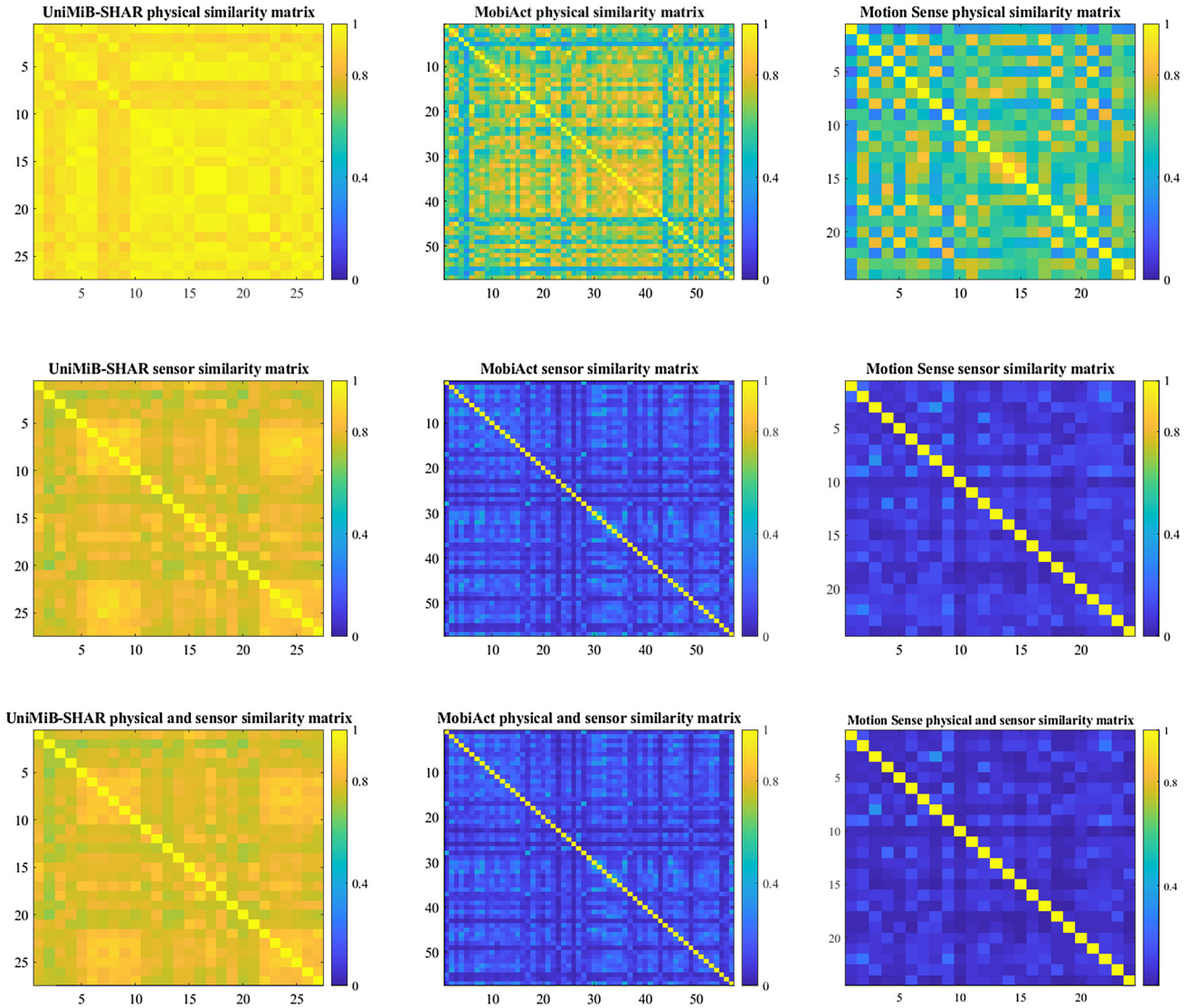

Fig. 2 Similarity matrices for physical, sensor, and their combinations of UniMiB SHAR, MobiAct and Motion Sense datasets

It is possible to state that PML models perform better with high similarities and small differences between subjects. In contrast, PDL models are less responsive to small differences and are highly impacted by the sample size.

It should be also pointed out that PDL outperforms PML models on Motion Sense because of the dataset size. In fact, Motion Sense is the dataset with the higher number of samples for each class (see Table 3). A larger amount of data permits to better fit deep learning models.

Finally, regardless of the model used (deep learning or machine learning), personalization seems to yield better results using a hybrid split and combined physical and sensor similarity.

5.2 RQ2: does deep learning outperform personalized machine and deep learning?

Table 5 summarizes the accuracy achieved from personalized deep learning (PDL, column PDL), personalized machine learning (PML, column PML), and traditional deep learning (DL, column DL). The values in column PDL are the best values of accuracy with respect to the value of the parameter m in Table 4, each with respect to the type of data split and type of similarity. The values of column PML are the same as in Table 4.

For sake of completeness, we also added a column (column ML) in which accuracy values obtained by applying traditional machine learning techniques are reported. Specifically, the values were computed using the AdaBoost classifier. The setup of the experiment is described in [10].

As in Table 4, also in Table 5 accuracies are grouped by dataset (column Dataset) and then by configuration (column Model): split type (subject-independent and hybrid; in Table 4 referred as SI and Hyb, respectively) and then similarity type (no similarity, physical, sensor, and physical in combination with sensor; in Table 4 referred as no sim, phy, sen, and phy+sen, respectively).

DL models outperform the other strategies in the most of the cases. PML overcomes DL only in the case of UniMiB SHAR dataset with hybrid models. Nevertheless, in UniMiB SHAR, DL strategies improve the overall accuracy in comparison with ML and PDL methods. This results is probably due to the strong similarity among the subjects in the dataset.

On average (last row of the table), DL models improve the performance of at least about 2%.

DL models show, in general, better results on MobiAct dataset with an accuracy equal to 92.62% with hybrid data split. In the case of subject-independent, 88.92% is achieved. This is an expected behavior because MobiAct is the largest dataset, which generally improves the classification capability.

For what concerns UniMiB SHAR, the best accuracy is achieved by the PML model with hybrid data split (84.87%). Nevertheless, in the subject-independent data split, DL still achieves the highest accuracy of 58.88%. Accuracy achieved with Motion Sense is on average 83.39%, by a margin from 4 and 10% with respect to the other techniques.

The results show that DL models are the most preferable in terms of robustness in comparison with PML, PDL, and ML techniques. Indeed, DL-based performance outperforms other methods performances even with different data split and different training datasets. The variability inter- and intra-subjects is overcome by DL. This result makes the DL method the one that achieves the highest generalization capability.

The comparison between DL and PDL methods leads to state that the training dataset size highly influences the algorithm's performance and normally large dataset are preferable. Indeed, the difference between PDL and DL methods is the training dataset's size.

In conclusion, DL algorithms are able to generalize user's differences and show very robust properties in terms of subject's variabilities. Thus, according to the results achieved in the three datasets, we can state that DL is confirmed as powerful method for human activity recognition.

Table 5 Experimental results—accuracy of personalized deep learning (PDL), personalized machine learning (PML), traditional deep learning (DL), and traditional machine learning (ML)

Dataset	Model	PDL	PML	DL	ML
MobiAct	SI-no sim	–	–	88.92	81.29
	SI-phy	86.08	81.62		
	SI-sen	80.14	83.45		
	SI-phy+sen	79.68	82.64		
	Hyb-no sim	–	–	92.62	83.73
	Hyb-phy	81.04	89.43		
	Hyb-sen	81.40	90.76		
	Hyb-phy+sen	88.17	90.90		
Average		82.75	86.46	**90.77**	82.51
UniMiB SHAR	SI-no sim	-	-	58.88	56.80
	SI-phy	35.42	57.39		
	SI-sen	42.83	57.00		
	SI-phy+sen	42.66	56.93		
	Hyb-no sim	–	–	69.72	61.66
	Hyb-phy	45.82	85.44		
	Hyb-sen	47.26	84.71		
	Hyb-phy+sen	46.77	84.87		
Average		43.46	**71.05**	64.30	59.23
Motion Sense	SI-no sim	–	–	81.03	72.48
	SI-phy	78.02	72.45		
	SI-sen	78.8	74.03		
	SI-phy+sen	79.00	73.85		
	Hyb-no sim	-	-	85.75	73.82
	Hyb-phy	80.17	77.76		
	Hyb-sen	80.38	78.06		
	Hyb-phy+sen	80.41	77.86		
Average		79.46	75.66	**83.39**	73.15
Total average		68.56	77.73	**79.49**	71.63

Values in bold correspond on average to the best accuracy obtained when varying the technique both with respect to individual datasets (rows Average) and to all the datasets (Total average row)

6 Conclusion

Continuous population's growing and aging are characterizing current and future eras. Life expectancy is estimated to grow longer and longer, particularly in high-income countries. Unfortunately, the increase in life expectancy is not always an added value for the individual since her/his last years of life are not always characterized by a good quality of life. Indeed, elderly may be affected from several age-related diseases, such as dementia, or they could simply require much efforts for their care from healthcare systems and their families. In this context, it is crucial to intervene with sustainable and long-term solutions.

Over the last decades, the progress in hardware and software technologies has encouraged the experimentation of several digital solutions for healthcare. Environmental and wearable devices have enabled the development of digital solutions for elderly monitoring, falls detection engines, and lifestyle monitoring, to name a few. In particular, wearable devices have gained the attention of the research's community. Their portability, efficiency, accuracy, and per-

vasiveness, make them attractive devices for researchers as well as for users.

Several techniques have been proposed for recognizing activities of daily living that exploit inertial signals collected by wearable devices. However, the lack of publicly available large datasets prevents the traditional algorithms from generalizing in real-world situations. In particular, algorithms struggle to generalize to new unseen users, because of the inter- and intra-variability among subjects.

In this work, we investigated personalized-based machine-learning and deep learning techniques, and compare their performance against traditional deep learning methods. The evaluation has been made relying on three of the most used datasets that include physical information about the subjects: MobiAct, UniMiB SHAR, and Motion Sense.

The achieved results show that traditional deep learning outperforms personalized techniques in most of the cases.

We can summarize the results achieved as follows.

- The size of the dataset seems not to affect the goodness of the classifier when subjects are dissimilar to each other.
- The more similar the subjects are to each other, the better the performance obtained using personalization.
- Personalization seems to provide better performance when applied to machine-learning techniques rather than deep learning techniques. This was found to be true for two out of three datasets (MobiAct and UniMiB SHAR). The third dataset (Motion Sense) performs better with personalization applied to deep learning techniques even though the difference in accuracy is very small (on the order of 2.8% in average).
- Regardless of the technique used (deep learning or machine learning), personalization seems to work better with a hybrid split and considering both personalization with sensors and with physical features
- Deep learning seems to provide better performance with respect to both personalized deep learning and machine-learning models.

The results obtained on the UniMiB SHAR dataset differ from the other datasets. This is due to the fact that the dataset contains subjects that are very similar to each other, both in terms of physical characteristics and in terms of signals.

Future investigations considering other datasets of inertial signals will allow to confirm what already obtained in this experimentation.

References

1. Abdallah ZS, Gaber MM, Srinivasan B, Krishnaswamy S (2015) Adaptive mobile activity recognition system with evolving data streams. Neurocomputing 150:304–317
2. Amrani H, Micucci D, Napoletano P (2021) Personalized models in human activity recognition using deep learning. In: Proceedings of the international conference on pattern recognition (ICPR)
3. Bishop CM (2006) Pattern recognition and machine learning. Springer, New York
4. Bulling A, Blanke U, Schiele B (2014) A tutorial on human activity recognition using body-worn inertial sensors. ACM Comput Surv (CSUR) 46(3):1–33
5. Chen K, Zhang D, Yao L, Guo B, Yu Z, Liu Y (2021) Deep learning for sensor-based human activity recognition: overview, challenges, and opportunities. ACM Comput Surv 54(4):1–40
6. Chen Y, Shen C (2017) Performance analysis of smartphone-sensor behavior for human activity recognition. IEEE Access 5:3095–3110
7. Ferrari A, Micucci D, Marco M, Napoletano P (2019) Hand-crafted features vs residual networks for human activities recognition using accelerometer. In: Proceedings of the IEEE international symposium on consumer technologies (ISCT)
8. Ferrari A, Micucci D, Marco M, Napoletano P (2021) Trends in human activity recognition using smartphones. J Reliab Intell Environ 7:189–213
9. Ferrari A, Micucci D, Mobilio M, Napoletano P (2019) Human activities recognition using accelerometer and gyroscope. In: Proceedings of the European conference on ambient intelligence (AmI)
10. Ferrari A, Micucci D, Mobilio M, Napoletano P (2020) On the personalization of classification models for human activity recognition. IEEE Access 8:32066–32079
11. Ferrari A, Micucci D, Mobilio M, Napoletano P (2020) Personalized deep learning in human activity recognition from inertial signals: a preliminary study on its effectiveness. In: Proceedings of the Italian workshop on artificial intelligence for an ageing society (AIxAS) co-located with international conference of the Italian association for artificial intelligence (AIxIA)
12. Friday NH, Al-garadi MA, Mujtaba G, Alo UR, Waqas A (2018) Deep learning fusion conceptual frameworks for complex human activity recognition using mobile and wearable sensors. In: Proceedings of the international conference on computing, mathematics and engineering technologies (iCoMET)
13. Garcia-Ceja E, Brena R (2015) Building personalized activity recognition models with scarce labeled data based on class similarities. In: Proceedings of the International conference on ubiquitous computing and ambient intelligence (UCAmI)
14. Garcia-Ceja E, Brena R (2016) Activity recognition using community data to complement small amounts of labeled instances. Sensors 16(6):877
15. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
16. Hong JH, Ramos J, Dey AK (2016) Toward personalized activity recognition systems with a semipopulation approach. IEEE Trans Hum Mach Syst 46(1):101–112

17. Igual R, Medrano C, Plaza I (2015) A comparison of public datasets for acceleration-based fall detection. Med Eng Phys 37(9):870–878
18. Lane ND, Xu Y, Lu H, Hu S, Choudhury T, Campbell AT, Zhao F (2011) Enabling large-scale human activity inference on smartphones using community similarity networks (csn). In: Proceedings of the international conference on ubiquitous computing (UbiComp)
19. Lara OD, Labrador MA et al (2013) A survey on human activity recognition using wearable sensors. IEEE Commun Surv Tutor 15(3):1192–1209
20. Lockhart JW, Weiss GM (2014) The benefits of personalized smartphone-based activity recognition models. In: Proceedings of the SIAM international conference on data mining (SDM)
21. Lockhart JW, Weiss GM (2014) Limitations with activity recognition methodology & data sets. In: Proceedings of the ACM international joint conference on pervasive and ubiquitous computing: adjunct publication (UbiComp)
22. Malekzadeh M, Clegg RG, Cavallaro A, Haddadi H (2018) Protecting sensory data against sensitive inferences. In: Proceedings of the workshop on privacy by design in distributed systems (W-P2DS)
23. Medrano C, Igual R, Plaza I, Castro M (2014) Detecting falls as novelties in acceleration patterns acquired with smartphones. PLoS One 9(4):e94811
24. Micucci D, Mobilio M, Napoletano P (2017) Unimib shar: a dataset for human activity recognition using acceleration data from smartphones. Appl Sci 7(10):1101
25. Organization WH (2015) World report on ageing and health. World Health Organization
26. Pires IM, Hussain F, Marques G, Garcia NM (2021) Comparison of machine learning techniques for the identification of human activities from inertial sensors available in a mobile device after the application of data imputation techniques. Comput Biol Med 135:104638
27. Reiss A, Stricker D (2013) Personalized mobile physical activity recognition. In: Proceeding of the IEEE international symposium on wearable computers (ISWC)
28. Rokni SA, Nourollahi M, Ghasemzadeh H (2018) Personalized human activity recognition using convolutional neural networks. In: Proceedings of the conference on artificial intelligence (AAAI)
29. Shen C, Chen Y, Yang G (2016) On motion-sensor behavior analysis for human-activity recognition via smartphones. In: Proceedings of the IEEE international conference on identity, security and behavior analysis (ISBA)
30. Siirtola P, Koskimäki H, Röning J (2019) Personalizing human activity recognition models using incremental learning. arXiv:1905.12628
31. Siirtola P, Röning J (2019) Incremental learning to personalize human activity recognition models: the importance of human AI collaboration. Sensors 19(23):5151
32. Sousa Lima W, Souto E, El-Khatib K, Jalali R, Gama J (2019) Human activity recognition using inertial sensors in a smartphone: an overview. Sensors 19(14):3213
33. Sztyler T, Stuckenschmidt H (2017) Online personalization of cross-subjects based activity recognition models on wearable devices. In: Proceedings of the IEEE international conference on pervasive computing and communications (PerCom)
34. Sztyler T, Stuckenschmidt H, Petrich W (2017) Position-aware activity recognition with wearable devices. Pervasive Mob Comput 38:281–295
35. Tapia EM, Intille SS, Haskell W, Larson K, Wright J, King A, Friedman R (2007) Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In: Proceeding of the IEEE international symposium on wearable computers (ISWC)
36. Vaizman Y, Ellis K, Lanckriet G (2017) Recognizing detailed human context in the wild from smartphones and smartwatches. IEEE Pervasive Comput 16(4):62–74
37. Vavoulas G, Chatzaki C, Malliotakis T, Pediaditis M, Tsiknakis M (2016) The mobiact dataset: Recognition of activities of daily living using smartphones. In: Proceedings of information and communication technologies for ageing well and e-Health (ICT4AgeingWell)
38. Viola P, Jones M (2004) Robust real-time face detection. Int J Comput Vis 57:137–154
39. Vo QV, Hoang MT, Choi D (2013) Personalization in mobile activity recognition system using k-medoids clustering algorithm. Int J Distrib Sens Netw 9(7):315841
40. Weiss GM, Lockhart JW (2012) The impact of personalization on smartphone-based activity recognition. In: Proceedings of the AAAI workshop on activity context representation: techniques and languages
41. Yu T, Chen J, Yan N, Liu X (2018) A multi-layer parallel lstm network for human activity recognition with smartphone sensors. In: Proceedings of the international conference on wireless communications and signal processing (WCSP)
42. Yu T, Zhuang Y, Mengshoel OJ, Yagan O (2016) Hybridizing personal and impersonal machine learning models for activity recognition on mobile devices. In: Proceedings of the EAI international conference on mobile computing, applications and services (MobiCASE)
43. Zhu R, Xiao Z, Li Y, Yang M, Tan Y, Zhou L, Lin S, Wen H (2019) Efficient human activity recognition solving the confusing activities via deep ensemble learning. IEEE Access 7:75490–75499
44. Zunino A, Cavazza J, Murino V (2017) Revisiting human action recognition: personalization vs. generalization. In: International conference on image analysis and processing. Springer, pp 469–480