SN Computer Science (2021) 2:133
https://doi.org/10.1007/s42979-021-00538-3

ORIGINAL RESEARCH

Daily Routine Recognition for Hearing Aid Personalization

Thomas Kuebert[1] · Henning Puder[1,2] · Heinz Koeppl[1]

Received: 11 September 2020 / Accepted: 22 February 2021 / Published online: 11 March 2021

Abstract

This work focuses on daily routine recognition to personalize the hearing aid (HA) configuration for each user. So far, there is only one public data set containing the data of two acceleration sensors taken under unconstrained real-life conditions of one person. Therefore, we create a realistic and extensive data set with seven subjects and a total length of 63449 min. For the recordings, the HA streams the acceleration and audio data to a mobile phone, where the user simultaneously annotates it. This builds the grounds for our comprehensive simulations, where we train a set of classifiers in an offline and online manner to analyze the model generalization abilities across subjects for high-level activities. To achieve this, we build a feature representation, which describes the recurring daily situations and environments well. For the offline classification, the deep neural network, multi-layer perceptron (MLP), and random forest (RF) trained in a person-dependent manner show the significantly best F-measure performance of 86.6%, 87.1%, and 87.3%, respectively. We confirm that for high-level activities the person-dependent model outperforms the independent one. In our online experiments, we personalize a model that was pretrained in a person-independent manner by daily updates. Thereby, multiple incremental learners and an online RF are tested. We demonstrate that MLP and RF improve the F-measure compared to the offline baselines.

Keywords Machine learning · Daily routine · Activity recognition · Hearing aid · Sensor fusion

Introduction

Hearing aids (HA) adapt their configuration to the incoming sounds based on a classification system. Therefore, in general, a class, e.g. speech in quiet, is linked to a predefined user-independent device setting [1]. Because this decision making is performed within seconds and the acoustic scene can rapidly change, the corresponding settings, e.g. frequency amplification, can often vary as well [2]. These frequent configuration modifications can be unpleasant while the HA wearer performs the same activity.

We aim to personalize these settings to the user's preferences and needs since the wearer's intention determines the ideal configuration in a certain situation. Assuming that the user does office work and the colleague besides him has a conversation with a visitor. Here, due to spatial proximity, the HA would decide that the user wants to listen to this conversation based on the short-term acoustic cues. However, the wearer's intention is to focus on his work. Hence, the audio information can be ambiguous, and we need to consider the user behavior over a longer period to deduce this kind of situations.

The goal is to provide a stable classification with less prediction ambiguity and a personalized HA configuration. Thus, we link the common, repetitive situations of the daily routine to a preferred setting. The slowly changing, periodic daily routine is a high-level activity, which is a composition of many low-level activities. Supporting this new concept of personalized HA scene adaption, we focus on the routine detection part. To design such a system, we built an acceleration (ACC) sensor in a HA to record over a longer period the motion patterns along with the acoustic features. Due to the resulting large data set, we develop an efficient processing scheme for the offline and online routine detection.

The paper is structured as follows. In Sect. 2, the related work on daily routine recognition (DRR), offline and online supervised approaches is described. In Sect. 3, the data set and routine annotations are introduced. In Sect. 4, the offline and online processing scheme of DRR is explained and

✉ Thomas Kuebert
kuebert@gsc.tu-darmstadt.de

[1] Technische Universität Darmstadt, Fraunhoferstraße 4, 64283 Darmstadt, Germany

[2] Sivantos GmbH, Henri-Dunant-Str. 100, 91058 Erlangen, Germany

applied to the routine data. Finally, the results are presented and conclusions are drawn in Sects. 5 and 6, respectively.

Related Work

The field of human activity recognition (HAR) has been intensively investigated by focusing on low-level activities. These studies showed good detection rates by finding suitable features for orientation, locomotion, transportation modalities, and gestures [3–6]. Thereby, researchers often used inertial measurement units (gyroscope, magneto- and accelerometer), where especially accelerometers were selected due to the low power consumption like in our case. In addition, it was demonstrated that for low-level activities the person-dependent model outperforms the independent one in various tasks [5]. Thus, in this contribution we address the question if this also holds for high-level activities. To assess this, we test different cross-validation schemes.

In audio research, lots of work was spent on good detection features for speech, own voice, and noise by their characteristics [7, 8]. The fusion of the HAR and audio fields was done in a few studies for mostly short-term activities such as in a workshop [9, 10]. We build on these results by applying the most suitable features for our DRR use case.

Whereas on the periodic daily routine, only a limited number of authors worked on these composition activities of low-level primitives by applying a topic model or co-occurrence statistics [11–13]. Especially the creation of data sets is extremely time-consuming and for accelerometers the TU Darmstadt set with one person exists [14]. Therefore, we address this gap by building a data set of multiple subjects with audio and ACC data for the HA preferred ear position and further analyze the routine data. In our previous work [15], we already showed the improved routine detection rates by combining audio and ACC features. Furthermore, we also applied our processing scheme on TU Darmstadt data set and showed a superior performance over the topic model [15]. For the offline recognition, lots of experiments with different classifiers such as decision trees or neural networks are performed for activity primitives mostly [5]. We continue these benchmark evaluations for the daily routine, which is expected to be more challenging due to the higher abstraction level that generates more variability.

Since the long-term activities might change over time and different routine compositions are carried out by the subjects, the online model personalization is also assessed to follow possible non-stationary behavior. Thereby, the classifiers are updated by adaptation of model parameters, ensemble methods or incremental updates [16, 17]. A survey on incremental learners stated a good tradeoff between the computational efficiency and performance by the linear support vector machine (SVM) with stochastic gradient descent updates, Gaussian Naïve Bayes and Online Random Forest (ORF) [18]. In addition, the popular neural networks are incremental learners by performing forward and backward passes on data chunks. Thus, we test all these algorithms and use as small considered multi-layer perceptron (MLP) network to keep the computational demands still feasible for a HA. The Gaussian mixture model (GMM) was often used in other audio or hearing aid studies for classification due to its computational efficiency and that is why we also apply it for comparison reasons [19, 20].

One study personalized the HAR model on inertial sensor data without an user interruption [21]. This is achieved by pretraining an user-independent model and performing incremental online updates with the own model predictions on unseen data. They used the Learn++ ensemble method, which adapts its model based on suitable-sized data chunks and tested three base classifiers. However, the user-independent model must be accurate enough that model personalization improves the recognition accuracy. Whereas in [22], the HAR model was personalized using an ORF. They updated the ensemble by adding new trees if sufficient samples have arrived and by deleting trees if their performance degraded in comparison to others with the new knowledge. These adding and forgetting mechanisms adapt to the new information. Both personalization studies worked on low-level activities with small data sets. We combine these update strategies in our own ensemble approach and cross-compare to existing incremental algorithms. Therefore, we intensively investigate the capabilities of these models to improve with their own predictions or true labels. Afterwards, the online evaluation assesses the performance either with the interleaved test-then-train, the so-called prequential, or holdout evaluation [23]. To compare with the offline baselines, we choose the holdout evaluation.

Data Set

In this paper, we propose to recognize the daily routine in an offline and online fashion for hearing aid personalization. Therefore, we consider the daily routine as a set of repetitive, common situations and environments among subjects. For this purpose, we create a realistic and extensive data set.

As this is a preliminary study to show the feasibility of the approach, the seven subjects are three females and four males with a low mean age of 29.3 ± 8.9 years for HA users. Hence, they are not representative for hearing aid customers, which are mostly in retirement age over 60 years old [2, 24]. Thus, it is expected that the younger people have a more active social lifestyle with more demanding hearing situations, which makes our task more challenging [25].

The goal for the subjects was to record the personal routine as long as possible (mean duration per day of 610.1 ± 166.7 minutes) over a longer period of time (mean number of 14.9 ± 3.4 days). During the total length of $N = 63449$ minutes, the Signia Nx hearing aid is worn on the ear and continuously streams the data via Bluetooth services to the mobile phone. The precomputed audio and raw acceleration features are ideally sampled at 2 Hz and 16 Hz, but sometimes due to transmission problems, the rate can be lower. The rates are optimized to have a stable transmission while keeping a good detection performance [5]. The variable rate of the data transmission leads to missing feature samples over time, which can have an influence on the classification performance [26]. Since our features are highly correlated over time from seconds up to minutes, the neighboring samples have similar information, i.e. the negative consequence of losing samples is reduced. Additionally, we design statistical features in section 4.1 that can deal with a variable number of feature samples. Thus, the daily routine detection is resilient to the missing feature problem.

Camera or raw audio recordings were considered but are not feasible over a long period of time and would be a privacy issue, especially in public environments. In contrast, our design is less obtrusive enabling the subjects to behave as natural as possible. Furthermore, the data timestamps and user annotations for the evaluation are generated in the mobile application. The users can report label errors, e.g. due to forgotten annotations or time offset, in the recording app for a later manual correction and shortly summarize their day.

The proposed routine classes have different hearing demands and are listed in Table 1. Starting at the top of the list, the transportation routine accounts for all modalities such as car or bus, going from A to B. While the physical activity stands for high-intensity routines like sport exercises or manual work, for instance. On the contrary, the basics group represents low-intensity activities and is inspired by the activities of daily living (ADL) concept [27], which represents the fundamental functions of living like eating or hygiene. Further activities, such as office work or reading a newspaper, are included as well.

Table 1 List of routine classes and corresponding activities

Routine	Activities
Transportation	Commuting, train, car, bus, plane, location change
Physical (Activity)	Exercises, sport, manual work
Basics	Hygiene, dressing, resting, eating, preparing food, housekeeping, office work
Social (Interaction)	Family, friends, conversations, partying, play music, singing, call
(Focused) Listening	Music, cinema, theater, concert, lecture, TV, media

The next two routine classes are influenced by the so-called common sound scenarios and are the most difficult situations for the hearing-impaired people [28]. The social (interaction) routine is the most crucial to participate in life during conversations in various environments. Likewise, the (focused) listening routine is another fundamental function for the hearing to receive information from media or joy from music. These two hearing functions are sometimes determined by the intention of the wearer in the situation. That is why the user should select the intended dominant routine, i.e. in a conversation during a car ride, the dominant routine would be social. Hence, the classes are not mutually exclusive, which may be a possible source of confusion for the classifier and may result in a lower recognition rate. But we assume, that the situational intention changes the motion behavior allowing us their detection.

The introduced classes correspond to different hearing needs, which require specific signal processing settings. A few non-exhaustive examples are mentioned to gain a better understanding of the routine class goals. In a listening or social situation, it is often required to focus on a target speaker, where directional hearing is beneficial. Whereas, in basics, transportation, or physical class an omni-directional setting helps to keep the situation awareness and monitor if someone approaches the HA user. In a car transport scene, a typical low-frequent noise is present, that creates the need for noise reduction measures.

A typical example day of our data set from one subject is shown in Fig. 1. The overall main activity routine is basics, which mostly consists of sitting at the desk and working on a computer with smaller interruptions, such as coffee breaks. Usually, during lunch break the nearby canteen is visited by foot with a loud babble background noise. Within the working day, some meetings plus general conversations are included in the listening or social routine. Furthermore, the main mode of transport is the bicycle or car for commuting. Typical evening routines are meeting friends as social class, watching TV as listening class, dancing or going to the fitness center as physical class. Five subjects had the described office work routine containing lots of repetitive situations and environments. Two subjects followed a less recurring schedule and had more free time activities.

Furthermore, the different personalities and routines affect the prior class distribution as shown in Fig. 2, i.e. some tend to be more talkative and others more a good listener [29]. Thus, the class imbalance varies across the subjects and the online personalization addresses this issue by adapting the classification models. Additionally, the representative features for each class play an important role in an imbalanced classification problem to separate the classes well [30]. That is why, we choose

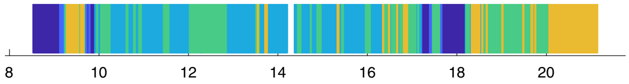


Fig. 1 Example day of one subject [hh]

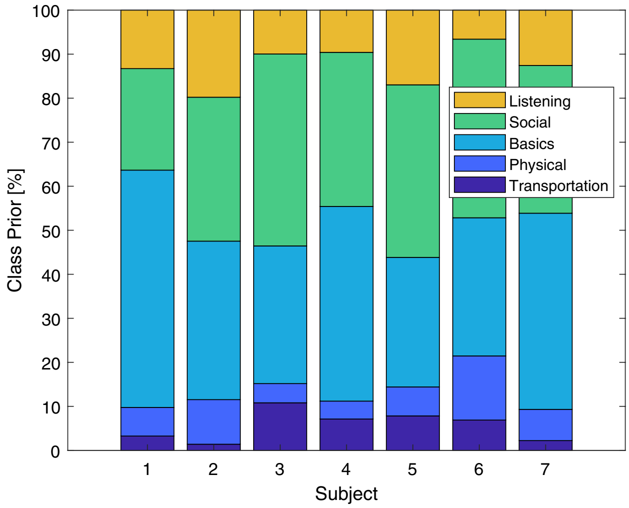


Fig. 2 Class prior per subject

the recorded features in the data set to be informative and discriminative for these classes.

Other potential public-available data sets are the Opportunity or TU Darmstadt set [14, 31]. Opportunity concentrates on ADL with lots of primitive activities in a sensor-rich and constrained environment. Whereas, the TU Darmstadt set contains the real life of the first author in an open setting by two ACC sensors. They focused on the daily routine during working days in office characterized by a very repetitive structure [15]. In contrast, our set contains the real life of 7 subjects and is a more realistic with unconstrained environments and activities. Since we deal with hearing aids, the preferred sensor location is the ear position on the head, which is not the case for the other two data sets. In addition, these previous studies did not use rich audio features as we do. Hence, there was the need to construct this realistic data set with multiple subjects.

Approach

In this section, the processing scheme for DRR shown in Fig. 3 is explained. First, the feature representation is built. Subsequently, the supervised learning scheme is applied in an offline and online fashion using various classifiers. Finally, the evaluation determines their performance.

Features

In the following, details to the features and their processing schematic, shown in the dashed block of Fig. 3, are given.

The features are built to distinguish the classes by representing the routine behavior and environments well. Their space can be partitioned in two independent inputs: ACC and audio. The raw acceleration is measured at a rate of 16 Hz and the precomputed audio features have a rate of 2 Hz. To fuse the inputs on the same time grid, the raw ACC is converted to features on an activity primitive level and then, a statistical representation is built on a routine activity level.

The rigid body model tells that the measured triaxial **acceleration** signal $\mathbf{a}_{\mathrm{mes}}$ is ideally only composed of gravitational $\mathbf{g}$, rotational, which splits in radial $\mathbf{a}_{\mathrm{R}}$ and tangential $\mathbf{a}_{\mathrm{T}}$, and linear $\mathbf{a}_{\mathrm{lin}}$ components:

$$\mathbf{a}_{\mathrm{mes}} = \mathbf{g} + \mathbf{a}_{\mathrm{R}} + \mathbf{a}_{\mathrm{T}} + \mathbf{a}_{\mathrm{lin}} \text{ in } [g], \qquad (1)$$

where all quantities are expressed in the sensor coordinate system and multiples of the earth gravity $g = 9.81\frac{m}{s^2}$ [32]. The vector $\mathbf{g}$ is only dependent on the sensor orientation, which is chained to the head and body orientation. If no motion is present, the gravity is directly given. The orientation is a key identifier to differentiate some scenes [3]. For example, in our case, sitting during office work and laying down in a workout can be distinguished. But in case of motion, the mean is a typical estimator for gravity [33].

The radial and tangential ACC are given by the cross product of sensor position vector $\mathbf{r}$ with angular ACC $\boldsymbol{\alpha}$ and velocity vectors $\boldsymbol{\omega}$:

$$\mathbf{a}_{\mathrm{T}} = \boldsymbol{\alpha} \times \mathbf{r} \quad \text{and} \quad \mathbf{a}_{\mathrm{R}} = \boldsymbol{\omega} \times \boldsymbol{\omega} \times \mathbf{r}. \qquad (2)$$

Both quantities are orthogonal to each other and that is why the correlation between axes gives clues about head

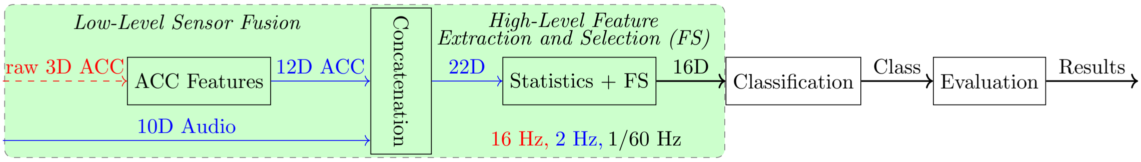


Fig. 3 The processing scheme of the audio and acceleration (ACC) inputs to recognize the daily routine, where the dashed block corresponds to the feature Subsect. 4.1

rotations. This allows us to detect conversational or listening gestures such as head shaking or nodding [34].

Furthermore, the linear motion dominates in comparison to head or body rotations, since the resulting amplitude is far stronger. That is why periodic movements, such as walking or jogging, produce a high output and often a variance feature is calculated to distinguish these from being stationary or sitting by the motion strength [6]. In addition, the mean crossing rate (MCR) informs about the motion frequency by counting the number of times the signal crosses the mean value, which is especially triggered by periodic movement. In total the four measures - mean, axes correlation, variance, and MCR - are extracted of the 3D ACC vector, which gives 12 dimensions. This is done over a sliding window of 1 second with 50 percent overlap, which demonstrated in other studies a good performance for the detection of activity primitives [35]. Conveniently, this choice matches the rate of audio features and fuse them on the same time grid.

For the **audio** signals, a set of 10 precomputed HA features is selected, since they describe well various environmental, music, and speech characteristics, which are helpful to detect the routine classes. The HA transforms the time signal to the frequency domain to compute these features. Thereby, the first band is from 0 to 125 Hz and the remaining 47 channels c have a width of 250 Hz up to 12 kHz. The listed features are grouped by the main detection property.

- The own voice activation takes advantage of the acoustic path from the mouth to the HA microphones [36]. It can distinguish between social and listening situations.
- The auto-correlation value of the actual sample and one a few milliseconds ago describes the tonality of music that can differ social or listening from other classes [7].
- Whereas, the wind activity helps to detect outdoor situations, but fast head rotations or movements can also trigger this feature due to the resulting airflow. It uses the non-existing correlation between the front and rear microphone signals.
- The maximum level

  $$l_{\max} = \max_{c} \; l_c \tag{3}$$

  of all bands $c \in 1, 2, \dots, 48$ gives clues about the loudness of the environment and demonstrated a good performance for various audio classification tasks [37].
- The spectral centroid (SC) of noise floor (NF), NF of low- and mid-frequency bands,

  $$SC_{NF} = \frac{\sum_{c=1}^{8} l_c \cdot c}{\sum_{c=1}^{8} l_c}, \tag{4}$$

  $$NF_{\text{Low}} = (\log_2 l_1 + \log_2 l_2)/2 \quad \text{and} \tag{5}$$

  $$NF_{\text{Mid}} = \log_2 \sum_{c=1}^{6} \frac{l_c \cdot w_c}{12} \text{ with } \mathbf{w} = [1\,2\,3\,3\,2\,1]^{\mathrm{T}} \tag{6}$$

  are good detectors for motorized modes of transportation, which produce a low-frequent noise [1].
- Three characteristic speech features are the average difference between level and noise floor, 4 Hertz modulation and onsets [1, 8].

The 22 audio and ACC low-level features are summarized in Table 2 and out of them we build the high-level routine representation. Therefore, they are segmented in non-overlapping one-minute frames to balance between fast audio (seconds) and slow activity (minutes) changes [38, 39]. This window length already showed a good performance in our previous work on TU Darmstadt and our former set [15]. Afterwards, the **statistical** quantities-mean, variance (var), and mean crossing rate-are computed for all features and frames [6]. This summarizes the information about gestures and low-level activities, e.g. the frequency of head rotations or strength of motion, and audio, e.g. changes in loudness levels or own voice activation, on a routine level. Thus, for example, the level of activity can distinguish the basics and physical routine. Whereas, the strength of speech properties or occurrence of low-frequent noise can separate the transportation and social routine.

To sum up, out of 22 low-level inputs three measures are extracted and 66 high-level features are returned, which are transformed to have zero mean and unit variance. Afterwards, we apply **feature selection** (FS) methods for the finding an optimal subset of features for the DRR [40]. Therefore, we first preselect a subset of 30 features with the minimal-redundancy-maximal-relevance criteria [41]. Then, we use the computational demanding wrapper-based approach of sequential feature selection (SFS) on the subset

Table 2 Summary of low-level features

Input	Methods
Acceleration (12D)	Mean, variance, mean crossing rate (axis-wise), axes correlation (between two axes)
Audio (10D)	Own voice activation, temporal auto-correlation, maximum level, spectral centroid, low- and mid-frequency noise floor, average difference, wind, 4 Hz modulation, onset detection

in a feasible amount of time. After that, our final feature representation contains 16 dimensions for routine recognition.

Classification

With the found ACC and audio features, we classify the routine behavior and environments in an offline and online manner. Therefore, a set of learners is selected for the evaluation, which are computationally feasible to use in a HA. For the **offline** classification, we perform batch learning on the entire training data and apply the fixed model on the unknown test set for the evaluation. We compare the recognition performance of the following classifiers:

- deep neural network (DNN) iteratively trains many parameters with three hidden layers consisting of 100 neurons per layer to avoid overfitting within the complex network an early stopping criteria and L2-regularization are applied,
- random forest (RF) builds an ensemble of 20 decision trees using randomization by bootstrapping samples for each tree and a random feature selection per binary split,
- multi-layer perceptron (MLP) iteratively trains a non-linear decision boundary with 100 hidden neurons,
- k-nearest neighbor (kNN) predicts the class of the nearest neighbor by finding the smallest Euclidean distance between the training examples and test sample,
- Gaussian mixture model (GMM) fits a mixture model of 2 components per class with a diagonal covariance and predicts using a maximum a posteriori (MAP) criteria,
- Naïve Bayes (NB) fits one Gaussian likelihood density per feature and class to decide per MAP criterion, and
- linear SVM parameterizes a hyperplane per class for the one-vs-all classification.

For the **online** classification, the initial model is trained on all known subjects and then personalized on the unknown test person P by daily updates. The initial training is performed in a leave-one-person-out (LOPO) manner and thus, it is called LOPO model. Thereby, the online personalization updates the classifiers based on the data of the new day in two ways either with the true labels (“**true update**”) or the own predictions (“**pred update**”). For the true update, we assume the user annotates the new data, for example, in a smartphone. Whereas, for the pred update, the current model predicts the labels of the new day and uses these for the training. Hence, we analyze if the classifier can self-improve over time without a necessary user-feedback.

Unlike in the offline phase with batch training, all classifiers are trained in data chunks, where the first one consists of all known subjects and the remaining ones are the daily updates of the new test person P. An one-day adaptation interval is chosen, since the daily routine activities are conducted over a time frame of minutes to hours. Thus, we ensure a broader data variability of multiple present classes for each daily adaptation, which should ease the model generalization. In doing so, we also want to imitate the behavior of a mobile system in real life, where the updates take place, for example, in a smartphone and only the adapted parameters are transferred to the HA. This is more computational- and energy-efficient than updates in shorter intervals. Therefore, we use the so-called partial or incremental fits of the MLP, NB, and SVM classifiers [18, 42], i.e. the parameters of the neurons, Gaussian densities, and hyperplanes are updated iteratively. For the GMM, the adaptation of the mean parameter is used since this is very efficient and changing the covariance showed only a minor improvement [43].

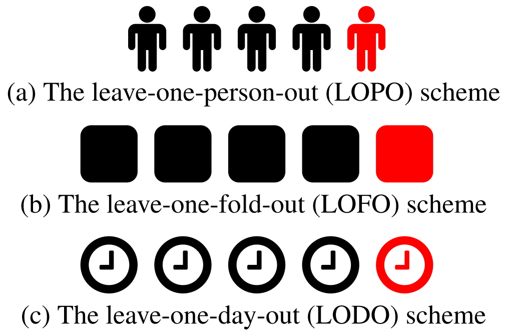
In our own ensemble method, we combine the Learn++ and ORF approach of [21, 22] by implementing an extended RF with an online mechanism, which adds and deletes ensemble trees. This adapts the model to possible instationary behavior or interpersonal differences. The online RF is constructed as follows the initial model is trained with 10 trees and for each daily update, we train two additional RF trees. This adapts to recurring daily behavior like in weekends. The fixed baseline ensembles train 20 trees to have the same total number of learners. The forgetting mechanism checks the individual accuracy of all ensemble trees and assumes the rates are Gaussian distributed. If a tree is worse than minus two times the standard deviation of mean performance, this tree is deleted. Furthermore, the online algorithms are compared to four baselines:

- the person-independent, initially-fitted LOPO model is called: “**fit(LOPO)**”,
- the person-dependent model that is only trained on the test person’s data: “**fit(P)**”,
- the combination of both personal and LOPO data in one model fit: “**fit(LOPO+P)**” and
- the person-independent classifier that is fine-tuned by the personal data: “**fit(LOPO)+adapt(P)**”.

All experiments are done in MATLAB R2019b in conjunction with classifiers from the Python library scikit-learn 0.22.2 [42]. For all methods, the made changes from the default parameters are explicitly mentioned. We implemented the online RF, offline and online GMM classifiers in MATLAB with the fitting functionality of Python for the RF trees and GMM probability distribution.

Evaluation and Experimental Setup

Evaluating the **offline** classification, the data set is split up in k parts as shown in Fig. 4. Afterwards on $k-1$ (black) subsets, the training is performed and on the unseen k-th (red) set the predictions are made. This process is repeated


(a) The leave-one-person-out (LOPO) scheme

(b) The leave-one-fold-out (LOFO) scheme

(c) The leave-one-day-out (LODO) scheme

Fig. 4 Offline evaluation schemes groups the data set by person, random fold, or day

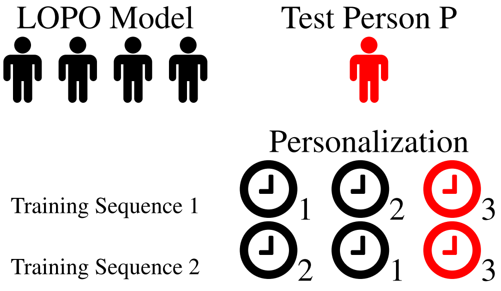


Fig. 5 Online evaluation scheme is a combination of a LOPO and LODO scheme

for all combinations, which gives a **cross-validation** (CV) scheme. Finally, the metrics are computed over all subsets.

The three applied CV schemes, **leave-one-person-out** (LOPO), **leave-one-fold-out** (LOFO), and **leave-one-day-out** (LODO), differ how they group the data set. Hence, LOPO splits person-wise, LOFO groups them in five random subsets of the same size and LODO makes one group per day. The LOPO scheme assesses the recognition rate of a person-independent model. On contrary to the LODO grouping, that is applied for a personalized training. As suggested in [44] for the LOFO scheme, there might exist a possible bias. Due to random split of the temporal data, neighboring samples can appear in different folds and are likely to be highly correlated. This results in over-optimistic recognition rates. We test if this bias is also present for high-level activities.

Evaluating the **online** classification, we use a combination of LOPO and LODO scheme shown in Fig. 5. First, the LOPO model is initially fitted on $k-1$ known subjects and the unseen subject k is used for online personalization. Therefore, the daily updates are performed on the training days and the recognition rate is reported on the fixed test day for each step. To further analyze the influence of training day order, we randomly permutate the training sequence as seen in Fig. 5. For the example of three days, the two possible training sequences 🕒$_1$, 🕒$_2$ and 🕒$_2$, 🕒$_1$ are displayed with a fixed test day 🕒$_3$. These training days 🕒$_1$ and 🕒$_2$ are the personal data of P, which is used to train three baselines. Of course, the performance metric is always estimated on the fixed unseen test day 🕒$_3$ and is averaged over all repetitions. Again, as in the LODO scheme, all day permutations are simulated, and the results averaged over all combinations. Afterwards, the online simulation is repeated for all subjects and the final performance is averaged over all outcomes. As it is typical in activity recognition, the measures,

- the confusion matrix summarized by four events: true positive (TP), true negative (TN), false positive (FP), and false negative (FN),
- accuracy $\frac{TP+TN}{TP+TN+FP+FN}$, and
- F_1-measure as harmonic mean of recall $\frac{TP}{TP+FN}$ and precision $\frac{TP}{TP+FP}$,

are applied [6]. We use the class-averaged F_1-measure and not the weighted version, since the data set has a strong class imbalance shown in Fig. 2 and the overall weighted performance would be dominated by the majority classes. That is why the reported rates are expected to be lower.

To compare the significance of two classifier results based on the performances of the CV folds, the Wilcoxon's signed-rank test is used, which as a non-parametric hypothesis test does not make a distribution assumption on the results [45]. Therefore, each result per fold is considered as a trial and the performance differences of the two classifiers are computed. The absolute values of these differences are ranked and for each classifier these ranks are summed, on which a classifier won the comparison. The lower sum is compared to a critical value and if it is lower, the null hypothesis is rejected at a confidence level of 5% that the performance of these classifiers is no different. Thus, one classifier is significantly better than the other.

Results and Discussion

In this section, the results of various classifiers are split up in an offline and online evaluation. Firstly, the offline outcome of different cross-validation schemes is presented and analyzed. Secondly, the online classification performance is compared and the interday variability is assessed.

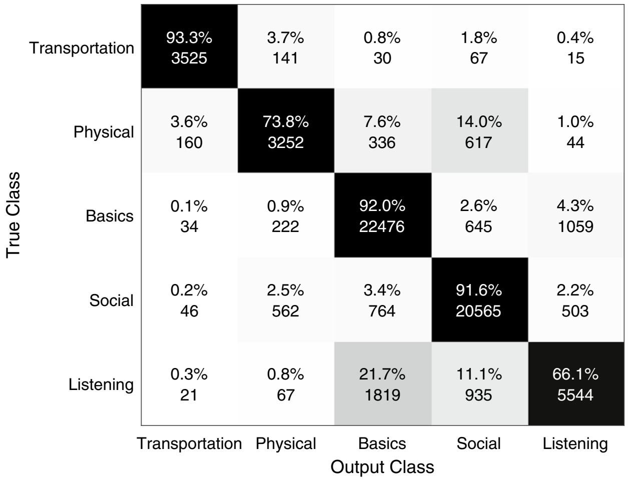
Offline Results

In the offline experiments, we analyze the F_1 performance of all classifiers based on three cross-validation schemes: LOPO, LOFO, and LODO. Thereby, we assess the person-dependent and -independent classification rate and look for a possible bias between the CV schemes. The comparison of various classifiers in the Table 3 states the very good F_1 results of the DNN and MLP network (84.3% and 84.6%) or the ensemble approach RF (84.1%) for the LOPO CV. The Wilcoxon hypothesis test shows that these three classifiers are significantly better than all others, but not to each other. The remaining learners also perform well in a close margin about 6% worse except the both density-based classifiers NB and GMM, which are more negatively affected by the class imbalance [46]. Further reasons for the ranking are that the linear decision boundary of the SVM is not complex enough to separate all classes. Thus, the MLP with a non-linear boundary distinguishes better between the classes. The lazy kNN classifier learns by example and does not generalize well over unseen data of different users [47]. The complexer DNN does not perform better than the MLP network, since the available amount of data is too less for the higher number of DNN parameters.

As mentioned in [44], the temporal correlation between consecutive samples in different folds boosts the F_1 results of LOFO over LOPO CV in Table 3 in the interval of 0.7% to 6.0%. This bias is smaller for the DNN and MLP network (4.2%) than for the ensemble method RF (5.2%), but worse for instance-based classifiers like kNN (6.0%). The parametric density estimation of NB (0.7%) and GMM (0.9%) is less affected by the temporal sample correlation, since the parameter updates are aggregated over the whole training data. This is also the case for the SVM parameterization of linear hyperplane and further non-tested classifiers, which follow the same learning principle.

Furthermore, we confirm the previous literature results [5, 22], where the person-dependent model with (LODO) CV scheme performs better than the independent one. This holds not only for low-level activities, but it is also valid for the high-level routine activities. The classifier ranking remains the similar to the LOPO case with smaller deviations.

Additionally, the detailed results of RF are presented by the confusion matrix in Fig. 6, where the class-wise recall is shown in the rows. Obviously, three of five classes are very well detected over 90% of recall and they contribute as majority to the high overall accuracy of 87.3%. The biggest confusion stems from the listening class with basics (21.7%) and social (11.1%). This makes sense due to the close relation between listening and social, where class transitions happen quit often. Likewise, the difference between listening and basics is mainly detected due to different audio characteristics, but for some situations they could be similar. For example, there is a background conversation and the subject does not want to follow it. Thus, a possible source for the classifier confusion stems from this intention-based scenery. Furthermore, the bigger mismatch of 14% between physical and social happens, since both classes could also be simultaneously and then the user's intention decide. Here, specialized acceleration conversation or movement features could deduce the motion behavior and the situational intention more precisely. Additionally, we analyzed if less transmitted data, i.e the missing feature problem, correlates with wrongly predicted samples. Therefore, a histogram with the number of transmitted samples per segment window given the correct or wrong prediction outcome states that both distribution are nearly identical. Thus, the daily routine recognition on our statistical features is robust to the missing feature problem.



	Transportation	Physical	Basics	Social	Listening
Transportation	93.3% 3525	3.7% 141	0.8% 30	1.8% 67	0.4% 15
Physical	3.6% 160	73.8% 3252	7.6% 336	14.0% 617	1.0% 44
Basics	0.1% 34	0.9% 222	92.0% 22476	2.6% 645	4.3% 1059
Social	0.2% 46	2.5% 562	3.4% 764	91.6% 20565	2.2% 503
Listening	0.3% 21	0.8% 67	21.7% 1819	11.1% 935	66.1% 5544

Fig. 6 Confusion matrix of LOPO with RF

Online Results

In the online simulation, we assess a possible performance improvement to four baselines by the daily model updates of the initially person-independent model. The online simulation results are dependent on the training sequences. One example is shown in Fig. 7, where the mean hold-out performance on the fixed test day is depicted over the various daily training updates with true or predicted labels and a confidence interval of one standard deviation (std). The

Table 3 Results of offline classifier F_1 performance [%]

CV	SVM	NB	MLP	kNN	GMM	RF	DNN
LOPO	77.3	70.8	84.6	78.4	68.1	84.1	84.3
LOFO	80.6	71.5	88.4	84.4	69.0	89.3	88.5
LODO	81.9	72.9	87.1	82.0	73.8	87.3	86.6

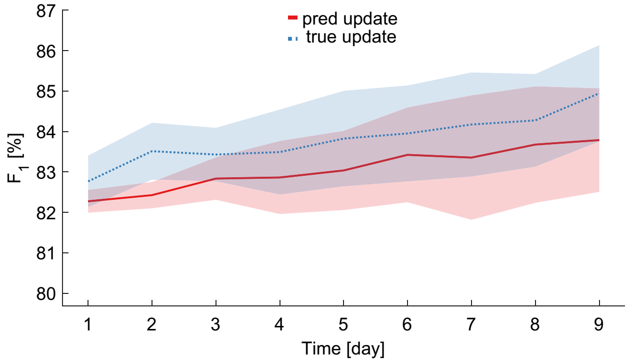


Fig. 7 Mean hold-out performance of multiple training sequences with a confidence interval of one standard deviation for the online MLP classifier

final performance is then taken as average over all sequences after the last training day. Obviously, the training updates improve the recognition outcomes for true and predicted labels. These results depend on how similar the training days so far and the test day are, which also contributes to the seen std and it slightly grows during both updates. For training updates with the own predictions, the results depend if the initial LOPO model is precise enough. Otherwise, the model cannot improve its performance, for that no example is shown. This confirms the previous study of [21].

Afterwards, the procedure is repeated for all combinations of test days and the performances are averaged. Then, the final F_1 results are obtained for all classifiers in Table 4 by averaging over all subjects. For MLP, NB, and RF only the true label updates are able to improve the results over the initial LOPO model performance by 0.6% to 0.8%. Therefore, the own model predictions are not on average reliable enough to improve the classifiers. The Wilcoxon hypothesis test shows that the MLP classifier is significantly better than all others on the true and predicted label updates. Again, the personalized model performs the best except for MLP, where the model improves more by having more data even from other subjects (fit(LOPO+P) and fit(LOPO)+adapt(P)). For the NB, the order of fitting does not matter, since it updates only its count statistics and densities. That is why the true label updates, fit(LOPO)+adapt(P) and fit(LOPO+P), have exactly the same performance of 66.7%. This is not the case for the RF, since three batch fits with personal data have a rounded value of 82.1%, but they have minor differences in the recognition rates. For the DNN, we denote the higher number of parameters hinders a model improvement during the online learning. Therefore, the simpler MLP classifier is in advantage and outperforms all other learning algorithms. The batch learning with personal data performs better than the online updates, since the interday variations are high and different activities are carried out on several days. Thus, learning with more present classes and activities generalizes better over unseen data.

Further analysis of the training sequences is depicted in Table 5, where the std of recognition rates over all training sequences is computed after the last update per person and is averaged over all subjects. Obviously, the GMM is stable with a very low std of 0.02%, since it only shifts the mean component. This update is independent of the order because the vector sum is associative. The MLP and DNN have a 0.54% and 0.58% smaller std for its own predicted labels than true ones, because the model makes consistent predictions, but not necessarily right ones. For the NB again, it updates only counts, where the order does not matter for the same result with the true labels. But in the case of the own predictions are used, the order changes the models over time. Thus, the predictions produce the different counts, which explains the slightly higher std of 0.11%. The RF std is similar to the MLP, which comes from the used model construction. Since the RF has inherent randomization by the data bootstrapping and random feature selection, the outputted trees are always different, which results in a slight std about 3%. The SVM linear hyperplane updates are more influenced

Table 4 Results of online classifiers F_1 performance [%]

Classifier	DNN	GMM	MLP	NB	RF	SVM
pred update	75.3	62.9	79.4	64.3	74.9	70.7
true update	75.0	63.3	81.0	66.7	78.9	70.8
fit(LOPO+P)	79.9	63.7	82.8	66.7	82.1	74.8
fit(LOPO)+adapt(P)	78.0	63.3	82.5	66.7	82.1	77.3
fit(P)	79.7	68.5	82.0	67.7	82.1	77.8
fit(LOPO)	77.0	63.4	80.4	66.1	78.1	73.2

Table 5 Standard deviation of training sequences (F_1 [%])

Classifier	DNN	GMM	MLP	NB	RF	SVM
pred update	3.79	0.02	2.20	0.11	3.34	8.77
true update	4.37	0.02	2.74	0.00	2.49	7.02

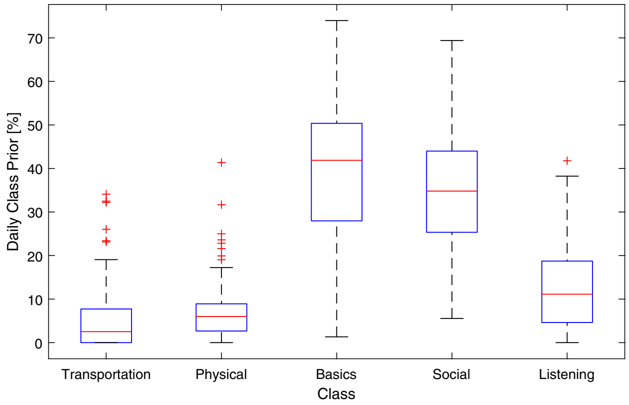


Fig. 8 The daily class prior of all subjects

with a std over 7% by the order of training days, since the interday variability of different classes is also high, i.e. the hyperplane shifts are too sensible with a small amount of data and do not generalize well.

To systematically analyze the interday class variability, we plot the distribution of the daily class prior in Fig. 8. Obviously, the two majority classes basics and social have the highest median values of 41.8% and 34.8%, but they also have the highest daily variability of 22.4% and 18.6% measured by the interquartile range. Whereas, only listening has a variability of 14.1% and the remaining two classes fall below 10%. These high differences in the class priors across days create the previously mentioned challenges for the detection algorithms and hinder an easy adaptation to this non-stationary process. According to [16], these changes are the so-called concept drift, which can occur sudden, reoccurring or incremental. In our case, the most relevant changes happen on a recurrent basis, since a weekend of lots free time activities strongly differs to a workday in office. We do not observe any sudden nor incremental drifts.

Conclusion and Outlook

In this work, we introduced as a first contribution a new real-life data set, which consists of 7 non-representative subjects for hearing aid (HA) wearers and a total length of 63449 minutes. The recorded acceleration and audio data describe the daily routine characteristics well due to our feature representation. On this basis, we perform two comprehensive comparisons for the offline and online routine classification. For the offline recognition, our second contribution confirms that the person-dependent model is superior to person-independent classifier. We further showed that the deep neural network, multi-layer perceptron (MLP), and random forest (RF) yielded the significantly best F-measure performance of 86.6%, 87.1%, and 87.3%. The remaining misclassified samples require a tailored motion representation to distinguish the intended behavior more precisely. In our online simulation, MLP and RF improved their F-measure performance by 0.6% and 0.8%, respectively, using the true labels compared to the baseline of the initially fitted model. Additionally, we analyzed the effect of the training sequence order and demonstrated a smaller influence of 2-3% at the F-measure rate for MLP and RF. The online analysis states our third contribution.

For future work, the routine detection performance is evaluated on a data set with HA users of a representative age range and over a longer period. Then, the personal dependency of these elderly users are assessed especially if their behavior patterns are different, other activities are performed, or a concept drift changes the routine distribution over time. However, we expect that these older subjects have a stronger and more repetitive routine, which should simplify their detection.

Acknowledgements We thank the Sivantos GmbH for funding the work of Thomas Kuebert and support, which made this contribution possible. Our special thanks go to the study participants for their dedication during the long recordings. Furthermore, the work of Thomas Kuebert is supported by the "Excellence Initiative" of the German Federal and State Governments and the Graduate School of Computational Engineering at Technische Universität Darmstadt.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Büchler MC. Algorithms for sound classification in hearing instruments. PhD thesis, ETH Zurich; 2002.
2. Dillon H. Hearing aids. Mumbai: Boomerang Press; 2012.
3. Zinnen A, Blanke U, Schiele B. An analysis of sensor-oriented vs. model-based activity recognition. In: IEEE International Symposium on Wearable Computers (ISWC); 2009.
4. Hemminki S, Nurmi P, Tarkoma S. Accelerometer-based transportation mode detection on smartphones. In: Proceedings of the

11th ACM conference on embedded networked sensor systems; 2013.
5. Lara OD, Labrador MA. A survey on human activity recognition using wearable sensors. IEEE Commun Surv Tutor. 2013;15(3):1192–209. https://doi.org/10.1109/SURV.2012.110112.00192.
6. Bulling A, Blanke U, Schiele B. A tutorial on human activity recognition using body-worn inertial sensors. ACM Comput Surv. 2014;46(3):33:1-33:33. https://doi.org/10.1145/2499621.
7. Peeters G. A large set of audio features for sound description (similarity and classification) in the cuidado project. Tech. rep., Institute for Research and Coordination in Acoustics/Music; 2004.
8. Scheirer E, Slaneyy M. Construction and evaluation of a robust multifeature speech/music discriminator. In: 1997 IEEE international conference on acoustics, speech, and signal processing, IEEE, vol 2; 1997. pp 1331–1334.
9. Lukowicz P, Ward JA, Junker H, Stäger M, Tröster G, Atrash A, Starner T. Recognizing workshop activity using body worn microphones and accelerometers. In: International conference on pervasive computing; 2004. pp 18–32.
10. Ward JA, Lukowicz P, Tröster G, Starner TE. Activity recognition of assembly tasks using body-worn microphones and accelerometers. IEEE Trans Pattern Anal Mach Intell. 2006;28(10):1553–67. https://doi.org/10.1109/TPAMI.2006.197.
11. Huynh T. Human activity recognition with wearable sensors. PhD thesis, TU Darmstadt; 2008.
12. Blanke U, Schiele B. Daily routine recognition through activity spotting. In: Lecture Notes in Computer Science, Springer Berlin Heidelberg; 2009. pp 192–206. https://doi.org/10.1007/978-3-642-01721-6_12.
13. Seiter JS. Topic models for activity discovery in daily life. PhD thesis, ETH Zurich; 2015. https://doi.org/10.3929/ethz-a-010483640.
14. Huynh T, Fritz M, Schiele B. Discovery of activity patterns using topic models. In: Proceedings of the 10th International Conference on Ubiquitous Computing, vol 8; 2008. pp 10–19, https://doi.org/10.1145/1409635.1409638.
15. Kuebert T, Puder H, Koeppl H. Daily routine recognition with visual interactive labeling by fusing acceleration and audio signals. In: 2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT); 2019. pp 1–6. https://doi.org/10.1109/ISSPIT47144.2019.9001791.
16. Ditzler G, Roveri M, Alippi C, Polikar R. Learning in nonstationary environments: a survey. IEEE Comput Intell Mag. 2015;10:12–25. https://doi.org/10.1109/mci.2015.2471196.
17. Laskov P, Gehl C, Krueger S, Mueller KR. Incremental support vector learning: analysis, implementation and applications. J Mach Learn Res. 2006;7:1909–36.
18. Losing V, Hammer B, Wersing H. Choosing the best algorithm for an incremental on-line learning task. In: European Symposium on Artificial Neural Networks; 2016.
19. Stowell D, Giannoulis D, Benetos E, Lagrange M, Plumbley MD. Detection and classification of acoustic scenes and events. IEEE Trans Multimed. 2015;17:1733–46.
20. Schädler MR, Kollmeier B. Separable spectro-temporal Gabor filter bank features: reducing the complexity of robust features for automatic speech recognition. New York: Acoustical Society of America; 2015.
21. Siirtola P, Koskimaki H, Roning J. Personalizing human activity recognition models using incremental learning. In: ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges (Belgium); 2018.
22. Sztyler T, Stuckenschmidt H. Online personalization of cross-subjects based activity recognition models on wearable devices. In: 2017 IEEE International Conference on Pervasive Computing and Communications (PerCom), IEEE; 2017.
23. Krawczyk B, Minku LL, Gama J, Stefanowski J, Woźniak M. Ensemble learning for data stream analysis: a survey. Inform Fusion. 2017;37:132–56.
24. Hasan SS, Chipara O, Wu YH. Evaluating auditory contexts and their impacts on hearing aid outcomes with mobile phones. In: Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare; 2014. pp 126–133.
25. Wu YH, Bentler RA. Do older adults have social lifestyles that place fewer demands on hearing? J Am Acad Audiol. 2012;23:697–711. https://doi.org/10.3766/jaaa.23.9.4.
26. Saar-Tsechansky M, Provost F. Handling missing values when applying classification models. J Mach Learn Res. 2007;8:1625–57.
27. Debes C, Merentitis A, Sukhanov S, Niessen M, Frangiadakis N, Bauer A. Monitoring activities of daily living in smart homes: understanding human behavior. IEEE Signal Process Mag. 2016;33:81–94. https://doi.org/10.1109/msp.2015.2503881.
28. Wolters F, Smeds K, Schmidt E, Christensen EK, Norup C. Common sound scenarios: a context-driven categorization of everyday sound environments for application in hearing-device research. J Am Acad Audiol. 2016;27:527–40. https://doi.org/10.3766/jaaa.15105.
29. John OP, Srivastava S. The big five trait taxonomy: History, measurement, and theoretical perspectives. Handb Personal Theory Res. 1999;2:102–38.
30. Chawla NV, Japkowicz N, Kotcz A. Editorial: Special issue on learning from imbalanced data sets. ACM SIGKDD Explor Newslett. 2004;6:1–6.
31. Chavarriagaa R, Saghaa H, Calatronib A, Digumartia ST, Tröster G, Millán J del R, Roggen D. The opportunity challenge: a benchmark database for on-body sensor-based activity recognition. Pattern Recogn Lett. 2013;34:2033–42.
32. Schopp P, Klingbeil L, Peters C, Buhmann A, Manoli Y. Sensor fusion algorithm and calibration for a gyroscope-free IMU. Procedia Chem. 2009;1(1):1323–6. https://doi.org/10.1016/j.proche.2009.07.330.
33. Karantonis DM, Narayanan MR, Mathie M, Lovell NH, Celler BG. Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. IEEE Trans Inform Technol Biomed. 2006;10(1):156–67. https://doi.org/10.1109/TITB.2005.856864.
34. Hale J, Ward JA, Buccheri F, Oliver D, de C Hamilton AF. Are you on my wavelength? Interpersonal coordination in naturalistic conversations. J Nonverbal Behav 2018.
35. Banos O, Galvez JM, Damas M, Pomares H, Rojas I. Window size impact in human activity recognition. Sensors. 2014;14:6474–99. https://doi.org/10.3390/s140406474.
36. Powers T, Froehlich M, Branda E, Weber J. Clinical study shows significant benefit of own voice processing. Tech. rep., Hearing Review; 2018.
37. McKinney MF, Breebaart J. Features for audio and music classification. Baltimore: Johns Hopkins University; 2003.
38. Huynh T, Blanke U, Schiele B. Scalable recognition of daily activities with wearable sensors. In: International Symposium on Location-and Context-Awareness, Springer; 2007. pp 50–67.
39. Kumar A, Raj B. Audio event detection using weakly labeled data. In: Proceedings of the 24th ACM international conference on Multimedia, Language Technologies Institute; 2016.
40. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3(Mar):1157–82.
41. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance,

and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27:1226–38.

42. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, *et al*. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
43. Reynolds DA, Rose RC. Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Trans Speech Audio Process. 1995;3(1):72–83.
44. Hammerla NY, Plötz T. Let's (not) stick together: pairwise similarity biases cross-validation in activity recognition. UBICOMP. 2015;10(1145/2750858):2807551.
45. Bifet A, de Francisci Morales G, Read J, Holmes G, Pfahringer B. Efficient online evaluation of big data stream classifiers. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining; 2015. pp 59–68.
46. Rennie JDM, Shih L, Teevan J, Karger DR. Tackling the poor assumptions of Naive Bayes text classifiers. In: Proceedings of the 20th international conference on machine learning (ICML-03), Artificial Intelligence Laboratory; Massachusetts Institute of Technology; Cambridge; 2003.
47. Zhang H. The optimality of naive Bayes. 2004;AA 1(2):3.