Health Technol. (2014) 4:273–287
DOI 10.1007/s12553-014-0084-9

ORIGINAL PAPER

Towards tailored physical activity health intervention: Predicting dropout participants

Xi Long · Marten Pijl · Steffen Pauws · Joyca Lacroix · Annelies H. C. Goris · Ronald M. Aarts

Received: 16 October 2013 / Accepted: 12 May 2014 / Published online: 28 May 2014

Abstract Physical activity is important for people's health. The physical activity intervention program reported here includes daily wearing of an activity monitor to provide people with insight into their activity behavior. The activity monitor consists of a triaxial accelerometer, where measured accelerations are transformed to a physical activity level (PAL). The PAL data quantifies the level of the daily physical activity and reflects the daily energy expenditure of the wearer. In the program, coaches provide e-mail based intervention to motivate participants to increase their activity step-by-step within 12 weeks. However, a significant portion of participants (∼41 %) failed to complete the program. This paper examines methods to predict participants who are at risk of dropping out of the program based on a classification task. This allows for a timely delivery of tailored interventions and motivating triggers to prevent stopping of the program. In particular, this paper proposes to combine the features extracted from participants' personal information, their behaviors during the use of the device, the observed PAL data and the features extracted from the process of predicting future PAL data to classify dropouts and non-dropouts every week. Experiment results show that a k-nearest-neighbor classifier achieved a dropout and a non-dropout prediction accuracy of 66.4 ± 13.8 % and 74.1 ± 7.3 %, respectively.

Keywords Physical activity level · Tailored intervention · Triaxial accelerometer · Dropout prediction

1 Introduction

Physical activity has large beneficial effects on people's mental and physical health [1]. The lack of engagement in a sufficient level of physical activity has been shown, e.g., to increase the risk of chronic disease, to deteriorate mental health, and to reduce the quality of sleep [2–4]. The increasing number of people with an inactive lifestyle may require the need for highly persuasive physical activity interventions to stimulate a healthier lifestyle and as such afford a better quality of sleep. Therefore, research into the development of effective physical activity promotion programs receives much attention nowadays [5, 6]. For participants of these programs, the provision of feedback about changes in physical activity behavior has proven to be highly important to stay motivated and to attain their goals [6, 7]. The use of wearable sensors is a promising solution to realize real-time monitoring and feedback provision of a person's physical activity behavior [8, 9]. Modern triaxial accelerometers are an inexpensive, effective, and feasible sensor that has often

X. Long (✉) · R. M. Aarts
Department of Electrical Engineering, Eindhoven University of Technology, 5612 AZ, Eindhoven, The Netherlands
e-mail: xi.long@philips.com,xi.long.ee@gmail.com

R. M. Aarts
e-mail: ronald.m.aarts@philips.com

X. Long · M. Pijl · S. Pauws · J. Lacroix · R. M. Aarts
Philips Research, 5656 AE, Eindhoven, The Netherlands

M. Pijl
e-mail: marten.pijl@philips.com

S. Pauws
e-mail: steffen.pauws@philips.com

J. Lacroix
e-mail: joyca.lacroix@philips.com

A. H. C. Goris
Philips DirectLife, 1096 BC, Amsterdam, The Netherlands
e-mail: annelies.goris@philips.com

been used to acquire activity information [8, 10, 11]. Previous studies have reported on the feasibility of wearing an accelerometry sensor for assessing daily energy expenditure by reliably estimating daily physical activity level (PAL) [12, 13]. PAL usually describes the energy consumption for physical activity as a fraction of the energy required to maintain basal metabolic functions [14].

1.1 Tailored activity intervention

In addition to feedback about behavior, the level of personalization of messages and program content is important for the effective support of behavior change [15]. One powerful way to personalize a program is through the incorporation of interaction with a human coach who has insight into the participant's behavior and who can provide support regarding specific motivational dips or barriers that the participant may encounter [16]. Nevertheless, human coaching support can be labor intensive, which limits the number of participants that a human coach can reach.

To make more efficient use of the sparse time of a coach, we propose to support the coach by identifying individuals that have the strongest need for coaching. This identification is based on models that can predict the intention of participants to drop out, which is assumed to correlate with a loss of motivation. The identification allows the coach to selectively direct attention to those individuals that are at risk of dropping out of the program, and provide them support to boost their motivation, reducing their risk of dropping out. Several studies have shown the beneficial effect of tailoring interventions to the motivational state of the receiver and providing support that satisfies their current motivational and behavioral needs compared to generic interventions [17, 18]. Identifying participants that run the risk of dropping out of the program in the future allows for a tailored intervention both in terms of timing and content of participant-coach interactions.

1.2 Physical activity intervention program

Many activity intervention programs have been studied in recent years [9, 16, 19]. In this study a 12-week physical activity intervention program was developed by Philips DirectLife (URL: www.directlife.philips.com) in order to increase the overall daily PAL of inactive people. The program is based on distributed user experiences such as activity monitor, computer (or website service), and e-mail, aiming at finding out the effectiveness on changing physical activity behavior in inactive people with different attitudes towards a change of physical activity.

Before the first intervention week (IW) of the program, participants were required to enter an assessment week (AW), in which they learned to use the provided device and completed personal characteristic data (e.g., age, gender, height, and body mass). In the AW, participants were asked to behave as they normally would and try not to increase their physical activity in order to offer a correct picture of their current activity level. After this week participants received a personalized activity plan and feedback, and they were gradually guided by a personal coach towards their final activity goal by means of daily-specified goals during the 12 weeks. For instance, several targets, e.g., minutes of high-intensity activity (such as running) and minutes of moderate-intensity activity (such as walking), were set by a coach during the program that aimed at motivating participants to achieve a higher level of physical activity.

In addition, the website provides participants with a historical overview of their performance and progress towards their goal, and the activity monitor allows for timely progress feedback. The participants can check their achieved personal activity score by observing the number of light-emitting diodes (LED's) that light up on the device (with a total number of 9 LED's). The personal activity score is the percentage of the measured physical activity level (PAL) to their daily PAL target, where no LED's lighting up corresponds to no activity and more than 5 LED's lighting up indicate that the target activity is reached. The PAL is obtained based on a built-in accelerometer, which will be explained later. The use of LED's enables the participants to assess their daily activity progress in an easy manner at any convenient time and adjust activity accordingly. More detailed description of the program can be found elsewhere [6, 7].

1.3 Description of dropout prediction problem

In regard to our real-life data, a significant portion of participants (∼41 %) failed to complete the program. The dropout prediction problem in a 12-week physical activity intervention program is explained in Fig. 1. We aim at predicting the participants who are likely to drop out in every future week based on the corresponding observed PAL data, recorded with a wearable accelerometry sensor (i.e., the activity monitor).

In the area of physical activity analysis, many studies have been performed with the use of body-worn accelerometers for different purposes such as sensor localization [20], posture and activity classification [21, 22], gait recognition [23], fall detection [24], respiration analysis [25], and epileptic seizure detection [26]. In addition, predicting future "events" has been investigated in other areas, such as driver action prediction [27], criminal behavior forecasting [28], and health management [29]. However, no studies have been found on predicting the dropout behavior in the physical activity domain. Our previous work [30] only focuses on predicting future data points rather than a specific event.

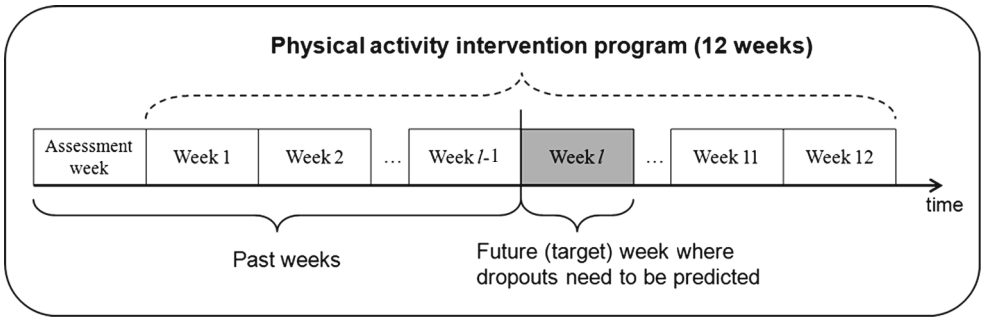

Fig. 1 Prediction of dropouts in week l ($l = 1, 2, ..., 11$) of the physical activity intervention program

In general, the existing methods for various prediction problems can be summarized into three categories: physical-model-based methods, qualitative-knowledge-based methods, and data-driven methods [31–33]. These methods usually work separately. In this study, we combined the objective measures, subjective data, and modeling information as the inputs of a dropout predictor, which will be described later.

1.4 Definition of dropouts and non-dropouts

In our study participants are either considered dropouts or non-dropouts (labeled as $dropout$ or $non\text{-}dropout$):

- Dropouts refer to participants who abandon the program prior to the end of the 12-week period. More specifically, a dropout is defined as a participant who does not record any activity during at least the last two weeks of the program. Note that there is no formal notification that a participant has quit the program; participants do not tend to tell their coach that they have stopped or will stop participation.
- Non-dropouts refer to participants who complete the 12-week program. Regardless of their PAL data throughout the program; there is at least recording of their activity during the AW and the last two weeks. For instance, during the period from week 1 to 10, if the participants stop the program for a few days and then rejoin later, they are considered non-dropouts.

Figure 2 shows the PAL data of a dropout and a non-dropout participant in the 12-week program. The identification of participants who are at a high risk of quitting within the program enables timely intervention, delivering coaching support to prevent them from dropping out.

1.5 Predictive classification

This paper explores the dropout prediction problem by means of predictively classifying of dropouts and non-dropouts in every future week. To classify them, some data elements obtained from past weeks should first be considered. These data elements are usually extracted from the raw data as "features". A set of traditional features can thus be derived from:

a) personal characteristic data,
b) observed PAL data in the past weeks,
c) behavioral data of participants during the use of the device in the past weeks (e.g., wearing days), and
d) data from a human coach in the past weeks (e.g., amount of targets reached).

In addition to them, we propose to extract novel features from a "PAL data prediction process". This process first uses a time-series model to fit the observed PAL data from the past weeks (modeling step) and then to predict the PAL data for the future week (prediction step) so that the features can be extracted from:

e) the model parameters (in the modeling step), and
f) the predicted PAL data (in the prediction step) in the future week.

To fit the observed PAL data and predict the future PAL data, an autoregressive integrated moving average-(ARIMA-) based method is employed, which has been well investigated for the same program in our previous

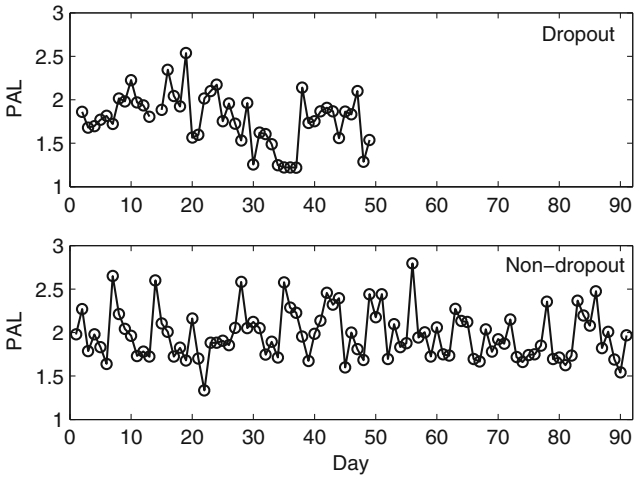

Fig. 2 Examples of the PAL data of a dropout and a non-dropout participant, where the dropout participant quitted the program on day 50 (week 7)

work [30]. These new features may contain information of a participant's intention or likelihood of dropping out in future weeks. More details about feature extraction will be provided in Section 4.1.

In terms of the number of weeks (one AW plus 12 IWs), the available database (see Section 2) is divided into 13 data sets. Here we only predict the dropouts in the first 11 IWs (from week 1 to 11) since a participant who recorded no activity during week 12 is not defined as a dropout. Note that participants only dropped out once in the program.

Several simple algorithms are considered to produce first results in classifying dropouts and non-dropouts such as a probability-based algorithm naive Bayes (NB), a regression-based algorithm linear discriminant (LD), a distance-based algorithm *k*-nearest-neighbor (KNN), and a hierarchical-based algorithm decision tree (DT). These classifiers usually assume that training samples are evenly distributed among different classes. However, the data set of each IW exhibits skewed class distribution, in which most of the samples (i.e., participants) fall into the majority *non-dropout* class and far fewer samples belong to the minority *dropout* class. It has been shown that the DT-based classifiers are sensitive to unbalanced class distribution [34] and therefore perform worse than the others in classifying an unbalanced data set [35–37]. This is because, for DT, the samples in a minority class presenting in leaves are often pruned. Moreover, the LD-based classifiers assume that features are normally distributed, but this assumption does not hold for some features in this study. NB and KNN are very simple classifiers which have been proven to be fast and with low computational complexity [38]. NB is not sensitive to irrelevant features and unbalanced data and it often performs well even if the assumption of feature independence does not hold [38, 40]. As a non-linear classifier, KNN does not require parametric assumptions and it is robust to noisy data [39, 41]. Therefore, we only compare NB and KNN in this study.

Frequently, classifiers induced from an unbalanced data set have a high accuracy for the majority class but an unacceptable accuracy for the minority class since they aimed at optimizing the overall accuracy [34, 42]. Many methods have been proposed to deal with this problem by resizing training sets, e.g., over-sampling the minority class samples [35] and/or under-sampling the majority class samples [43]. However, the removal of training sets may result in losing important information. On the other hand, random copies of minority samples may lead to over-fitting. A cost sensitive learning (CSL) method has been successfully developed to handle this unbalanced problem by assigning distinct misclassification costs to minority and majority classes [44]. In this paper, a weighted CSL (WCSL) method is proposed where the costs are adjusted by a weight.

Additionally, a genetic algorithm (GA) is expected to enhance the classification performance by maximizing the inter-class difference and minimizing the intra-class difference of features simultaneously [45]. Since GA will generate a large number of new features which might be mutually correlated, we employ principal component analysis (PCA) to reduce feature vector dimension and remove the correlation among those features. Both GA and PCA will be described in Section 4.

2 Data collection

2.1 Activity monitor

In the physical activity intervention program, to promote a healthier lifestyle by being more physically active, participants were provided with a wearable sensor: the Philips DirectLife activity monitor with a built-in triaxial accelerometer to measure the acceleration data of their activities performed throughout the day. As shown in Fig. 3, the monitor is a light-weight (12.5 g) and small-size ($3.2\times3.2\times0.5$ cm) portable device. It can be worn easily on the human body in a free-living environment (e.g., on the chest with a key cord, on the waist, or in the trouser pocket in an arbitrary orientation). The features of the device have been designed to enhance unobtrusiveness of wearing and to reduce the interference of the monitoring system with spontaneous activity behavior. During the monitoring period, the activity monitor was connected several times to a personal computer using Universal Serial Bus (USB) communication and the recorded data were uploaded, processed and stored using dedicated software.

2.2 Estimation of physical activity level

The output of the activity monitor is expressed as activity counts per minute (AC/m), which is the running time summations of absolute output values from the three axes of the accelerometer in the device. Consecutive counts were summed to arrive at counts per day, yielding an output of activity counts per day (AC/d) which has been proved to be linearly related to the PAL [8, 13]. A simple linear regression model can be used to accurately estimate the PAL values depending on AC/d, obtained with the activity monitor in free-living conditions [13]. However, the estimation of PAL may also correspond to the type of physical activity. Thus, models that account for the type of activity performed should offer a more accurate estimate of PAL. An improved model has been proposed to estimate PAL values using AC/d, corrected for the daily durations of six activity types (i.e., lying, sitting/standing, active standing, walking, running, and cycling) [46]. In that work, the six activity

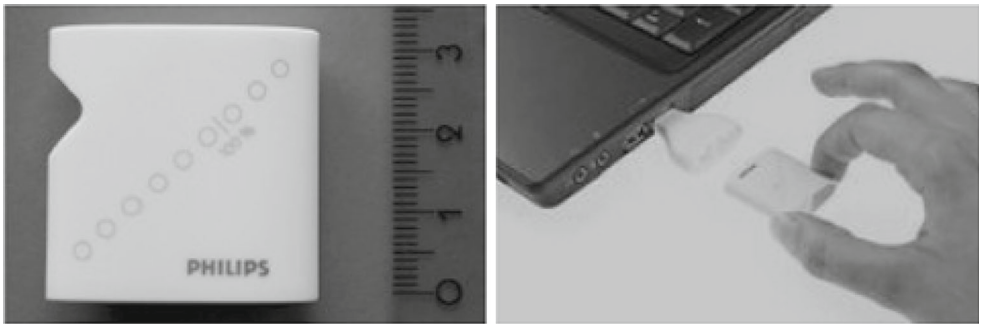

Fig. 3 The Philips DirectLife activity monitor (*Left*) and its connection to a personal computer via USB for uploading data (*Right*)

types performed during the day were identified based on the raw acceleration data using a classification tree algorithm. The model based on AC/d achieved a low standard error of estimate (SEE) of 7.3 % of the mean measured PAL. In this study, we adopted that model to compute PAL values.

2.3 Participants and database

The physical activity intervention program was primarily enrolled at different locations in the Netherlands throughout the year with a high participation in the months November and December 2008 and January 2009. In total 950 participants (587 males and 363 females), with an average age of 42.9 $\pm$ 9.8 years (ranging from 15 to 68) and body mass index (BMI) of 25.8 $\pm$ 6.2 kg/m^2, were recruited in the program. They were employees from two companies and most of them had one office function within a period of 13 weeks. Each participant took part in an AW and 12 IWs. As mentioned, participants learned to use the provided monitor and completed their personal characteristic data (i.e., age, gender, height, etc.) during the AW. No instruction about where and how to place the activity monitor was provided to participants so that they used the device according to their lifestyle circumstances. The PAL data recorded during the AW serves as a baseline measurement. In total, there are 91 days (13$\times$7) in which participants wore the device. Each day consists of a data point containing the PAL score accumulated over an entire 24-hour day of a participant in the program. Note that data were obtained anonymously and all participants provided an informed consent.

A daily PAL of between 1.2 and 2.5 is considered healthy for adults [47]. The lower bound refers to a sedentary level. PAL scores that are lower than this bound are treated as indicative of not wearing the device. Values above the upper bound correspond to an extremely high level of physical activity, e.g., endurance training potentially results in a PAL score as high as 4.5 [48]. We consider the very low ($<$1.2) and high ($>$5) PAL scores to be erroneous measurements and therefore treat these as missing data. Outliers and missing data points are generated by non-modeled mechanisms such as not wearing the activity monitor, a flat battery, monitor noise, or other disturbances. These data points are simply interpolated by using the overall mean value of the corresponding time series.

Besides the PAL data measured by the activity monitor, a real-life database is available in which various data elements were logged about each participant to track progress in his/her program. They include personal characteristic data, behavioral data, and data from the human coach as mentioned previously. Among which, the behavioral data include the device docking, website login, LED activation, daily wearing hours of device, etc. For instance, the daily wearing hours is estimated as the duration with AC/m larger than a given threshold.

3 PAL data prediction

Predicting future week's PAL data aims at further extracting features from the data fitting process (i.e., the modeling step) and from the predicted PAL data for dropout and non-dropout classification.

ARIMA-based methods have been widely used for time series analysis [49, 50]. In this study, a categorized-ARIMA (C-ARIMA) method is used to predict PAL data in each single future week based on the observed data. It has been reported that the C-ARIMA method achieved the PAL data prediction with a good performance in terms of prediction accuracy, model parsimony, computational load, and robustness compared with the traditional ARIMA models [30]. Here we briefly introduce this method, which consists of two steps: data fitting (or modeling) and data prediction. Note that only a subset of the database (227 out of 950 time series, 90 dropouts) without any missing values and outliers are kept for modeling because including them will lead to model misspecification and bad prediction performance. The remaining 723 series (301 dropouts) are used for model validation. To implement the PAL data prediction, we used the State Space Models (SSM) toolbox, developed by Peng and Aston [51], in Matlab R2012a (The Mathworks, Natick, MA).

3.1 Data fitting

Data fitting aims at constructing the most appropriate ARIMA model to fit the observed data from past weeks. A general ARIMA model can be structurally classified as the form of ARIMA$(p, d, q)(P, D, Q)^S$ [52], whereby the symbols are defined as follows: p the order of non-seasonal autoregressive (AR) terms, d the order of non-seasonal differencing, q the order of non-seasonal moving average (MA) terms; P the order of seasonal autoregressive (SAR) terms, D the order of seasonal differencing, Q the order of seasonal moving average (SMA) terms, and S the seasonal order. The C-ARIMA modeling consists of identification, model estimation and order selection, and diagnostic checking, which has been previously studied [30].

- To ensure that the model fits the observed PAL time series best, the stationarity, seasonality and trend of the observed data can be tentatively identified based on autocorrelation function (ACF) analysis [53].
- For model estimation, we apply a well-known recursive algorithm, Kalman filtering [54], to the observed PAL series.
- The orders can be selected using the Schwarz's Information Criterion (SIC) [55]. The best ARIMA models for the stationary, trend, and seasonal data have been found to have orders of $(1, d, 1)(0, D, 0)^7$, $(0, d, 1)(0, D, 0)^7$, and $(1, d, 0)(0, D, 1)^7$, respectively [30].
- Diagnostic checking is used to ensure that the fitting residuals for the time series are uncorrelated and normally distributed. This has been done in [30], indicating that the fitting models are adequate.

3.2 Data prediction

Regarding the C-ARIMA method used in this study, the most appropriate model is the one that best fits the pre-identified category to which this time series belongs (i.e., stationarity, trend, or seasonality). Afterwards, the chosen model should be used to predict future data. Based on our previous work [30], we obtained the multiple-step-ahead (from 1 to 7 steps) errors of predicting PAL data in terms of mean absolute percentage error (MAPE) for all target weeks (Fig. 4). A high precision in predicting PAL data allows for extracting features from these predicted PAL data with more adequate information in discriminating between dropouts and non-dropouts. In addition, the C-ARIMA method has been proved to be robust against noise and missing values [30].

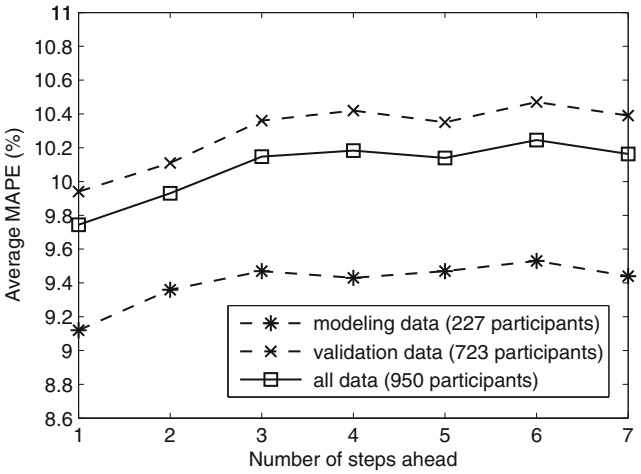

Fig. 4 Average mean absolute percentage errors (MAPEs) of 7-step-ahead data predictions using the C-ARIMA method described in [30]

4 Dropout and non-dropout classification

In order to allow for timely interventions to prevent future dropouts, they need to be identified as such before actually dropping out of the program. The dropout week is the week of the program corresponding to the dropout day which is defined as the first day on which no activity is recorded for the remainder of the program. As dropouts express the participants who have no recorded activity in the last two weeks of the 12-week program, the set of dropouts is divided into 11 separated sets of the IWs, among which, dropouts in week 11 have the first day of week 11 as their actual dropout day. The aim in this study is to classify each dropout in the week prior to their dropout week using known data only up to the dropout week. The non-dropouts remain present throughout the program. All the 950 participant samples in the database are included in this classification task.

4.1 Feature extraction

A number of features regarding different aspects of a participant (Section 1.5) are extracted from the database. They are grouped as 1) *personal*, 2) *assessment*, 3) *dynamic*, 4) *global*, and 5) *predicted* (or ARIMA) features, where the features in the former four groups are called "traditional features". The features obtained from the AW that act as a baseline are used in the classification of every other week. The five groups of features are described in the following.

Personal features are derived from the personal characteristic data provided by participants such as age, gender, height, BMI and program starting month.

Assessment features are computed over the duration of the assessment period, which typically lasts for a week,

although the exact length varies somewhat between participants. They provide information on the activity and behavior of the participant during the assessment period. Features on activity are the average and standard deviation of PAL scores, average of the logarithms of PAL scores, daily moderate-intensity and high-intensity PAL minutes. Features regarding behavior are the percentage of device wearing days, daily wearing hours, daily device docks, daily times of LED activation (or lighting up) on the device (indication of the transitions from physical inactivity to activity) and daily website logins.

Dynamic features are similar to the assessment features, but are extracted based on the observed PAL data in all the past weeks except the AW (i.e., the past IWs). Every dynamic feature is computed over a duration lasting a single week or over the past IWs, and the features extracted from every past IW are also included. The number of past IWs determines the number of dynamic features, which change over time by weeks. For instance, if we want to predict dropouts in week l ($l = 2, 3, \ldots, 11$), which also means the number of past IWs is $l - 1$, the dynamic feature set contains features extracted from the observed PAL data in week 1, week 2, …, and week $l - 1$. Here the number of features is $l - 1$ times that in the case of prediction dropouts in week 2. Note that there are no dynamic features when predicting dropouts in week 1 ($l = 1$). The dynamic features for activity and behavior are the same as for the assessment features, with the exception that the amount of targets reached and the average ratio of activity versus targets are considered as well.

Global features are computed over all the past IWs as shown in the following examples. In the time domain, the ratio of the highest to the lowest PAL indicates physical activity range of a participant. The corresponding distance in the time course describes the difference between the days with the highest and lowest PAL scores. As it is assumed that a PAL time series reflects weekly periodicity (from Monday to Sunday), then the features of averages and standard deviations of PAL scores on all the Mondays, Tuesdays, …, or Sundays should indicate average patterns of weekly-based activities of a participant. The maximal and minimal values among these averages show in which day(s) of the week a participant is the most active or inactive. In the frequency domain, the dominant frequency in the spectrum of an observed PAL data series is extracted. Frequency-domain entropy may help with distinction of the activity level of a participant with similar power intensity by comparing its periodicity. It is computed as the information entropy of the normalized power spectral density function of the input data without including the DC component. In addition, the amount of missing values and outliers in a PAL time series implies whether there was a device problem (e.g., a flat battery) during the use of an activity monitor for a participant or how active he/she was in wearing the monitor in the program.

Predicted (ARIMA) features are extracted from the ARIMA modeling process and the predicted PAL data in future week. In the modeling process, examples are AR and MA coefficients, data pattern type (i.e., seasonality, trend, or stationarity), SIC value and average fitting error (measured by MAPE). For the predicted PAL data, the features derived are the average and standard deviation of PAL scores, the highest and the lowest PAL scores and the ratio between the highest and the lowest scores. Here we use the C-ARIMA method that can achieve a low computation time and good accuracy in fitting the observed PAL data and in predicting the future PAL data (see Section 3).

4.2 Genetic algorithm

GA has proven effective in combining and selecting features in a classification problem which can help improve the classification performance [45]. The algorithm starts from a pool of potential solution candidates, known as the population. Each candidate consists of a fixed number of so-called chromosomes, which can represent, for instance, the inclusion of a certain feature, a numerical operator, etc. The candidates are altered and combined over a number of rounds in an iterative process. The aim of GA is to semi-randomly generate effective sets of chromosomes and then combine them into an increasingly optimal set of candidates over a number of rounds. A key feature of GA is the ability to search for globally optimal solutions, although it cannot be guaranteed that a global optimum is found. For more details about GA, we refer to [56].

4.3 Principal component analysis

PCA is a matrix conversion approach which represents a set of vectors (or components) in a new space with usually a lower dimension, where the vectors in the new space are mutually uncorrelated (or independent when the vectors are normally distributed) [57]. In other words, it highlights the similarities and differences of feature values. PCA not only reduces the feature vector dimension without much loss of information but also allows for removing the correlation among features expected to help improve the overall "quality" (or discriminative power) of the features as a result of redundancy and noise removal [11].

4.4 Classifier

NB - given a test sample u, the naive Bayesian classification function is expressed as

$$\mathbf{C}(u) = \arg\max_{c \in C} P(C = c) \prod_{i=1}^{m} p(F_i = f_i | C = c), \quad (1)$$

where C represents the class ($dropout$ or $non\text{-}dropout$), taking the value $c = c_{dropout}$ or $c = c_{non\text{-}dropout}$ in this study. For the given sample u, f_i is the value of feature F_i, where $i = 1, 2, \ldots, m$ with m features. $p(C = c)$ is the prior probability of the training data of the two classes. The conditional probability density function of the values of each feature F_i given a class can be obtained by fitting its distribution [58]. The multiplication holds when features $F_1, F_2, \ldots, F_m$ are mutually independent in Eq. 1. The equation also implies that the class with maximum conditional joint probability is selected. The NB classifier simply assumes that the features are mutually independent given the class.

KNN - in the nearest neighbor classification, the distance between a new sample and the nearest training sample is calculated. The KNN classifier makes an extension by taking the k nearest neighbors; it simply selects the class with majority votes. Given a test sample u, its k nearest neighbors $v_1, v_2, \ldots, v_k$ are found by calculating the Euclidean distances of the features between u and the other samples in the training set. Then a vote is conducted to assign the most common class to u. The class of u, denoted by $C(u)$ which takes the value $c = c_{dropout}$ or $c = c_{non\text{-}dropout}$, is determined by the function

$$\mathbf{C}(u) = \arg\max_{c \in C} \sum_{i=1}^{k} \delta(c, c(v_i)), \quad (2)$$

where δ is a function that $\delta(\alpha, \beta) = 1$ if $\alpha = \beta$ or $\delta(\alpha, \beta) = 0$ if $\alpha \neq \beta$, $c(v_i)$ is the class of v_i. The choice of k is important to the classification performance. It may be affected by the noisy neighbor points when choosing a too small value for k. On the other side, a large value for k is able to reduce the effect of noise but makes the classes less distinct. For small data sets, cross validation is usually used to determine the value of k, which will be explained later.

4.5 Weighted cost sensitive learning

During the decision making of classification, traditional learning methods assume a same misclassification cost for all classes. A CSL-based method assigns distinct costs to different classes with different numbers of samples in the training set. A high cost of misclassification should be assigned to the minority class while a low cost of misclassification should be assigned to the majority class for the situation of an unbalanced class distribution. The use of non-uniform error costs, defined by means of the class unbalance ratio (calculated by their prior probabilities) present in the data set, was proposed in [34]. Here the dropout misclassification cost $R_{dropout}$ and non-dropout misclassification cost $R_{non\text{-}dropout}$ for decision making are defined as

$$R_{dropout} = \frac{\omega}{p(C = c_{dropout})} \quad (3)$$

and

$$R_{non\text{-}dropout} = \frac{1 - \omega}{p(C = c_{non\text{-}dropout})}, \quad (4)$$

in which the weight ω is used to adjust the misclassification costs for the two classes. For instance, the costs only depend on the class unbalance ratio when $\omega = 0.5$. The classification by using this WCSL method can be optimized by examining ω for the classifiers on training data.

5 Experimental evaluation

5.1 Feature discriminative power

To evaluate the discriminative power of each feature, namely how good a feature can perform, in classifying dropouts and non-dropouts, a Hellinger distance metric [59] is employed. It is estimated by computing the amount of overlap between two probability density estimates, expressed as

$$D_H(p, q) = \sqrt{1 - \sum \sqrt{p(x)q(x)}}, \quad (5)$$

where $p(x)$ and $q(x)$ are the probability density estimates of the feature values given class $dropout$ and $non\text{-}dropout$, respectively, and $D_H \in [0, 1]$. In its most basic form, these density estimates can be computed by means of a normalized histogram with a fixed number of bins. In our study we empirically computed with 20 bins. A larger Hellinger distance reflects a higher discriminative power in separating the two classes.

5.2 Evaluation criteria

In a binary-class problem here, tp (true positive) and tn (true negative) refer to the number of dropout and non-dropout samples correctly classified, whereas fp (false positive) and fn (false negative) refer to the number of misclassified non-dropout and dropout samples, respectively. As there are 11 weeks' data under consideration, the database is divided into 11 data sets when performing classification. These data sets happen to be unbalanced. They contain a

majority of non-dropout samples (94.1 %) with a minority of dropout samples (5.9 %) on average for each target week. The overall accuracy, defined as (tp+tn)/(tp+fn+fp+tn), is not an appropriate evaluation criterion since it is strongly biased to prefer the majority class and is sensitive to skewed class distribution [60]. Thus, the classification accuracy for the two classes should be dealt with differently. To evaluate the classification performance for an unbalanced data set, we apply the *F-score* [61], defined as

$$F\text{-}score = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}, \quad (6)$$

where recall is tp/(tp+fn) and precision is tp/(tp+fp). This metric focuses more on the $dropout$ class by considering both recall and precision. Here the recall and precision are weighted equally, which implies the one may dominate the *F-score* rather than the other in an unbalanced data set. Actually, they can be weighted differently and their weights normally depend on which one is more emphasized in practical use. Therefore, it may not be guaranteed that the *F-score* in Eq. 4 is a perfectly suitable metric for evaluating classification performance although it is much better than overall accuracy. Moreover, for an unbalanced data set, the Receiver Operating Characteristic (ROC) curve has been widely used to assess a classifier's performance over the entire solution space instead of using a single metric [37]. A ROC curve is plotted as tp rate (sensitivity or recall) versus fp rate (one minus specificity), where sensitivity (in terms of dropout accuracy) is computed as fp/(fp+tn) and specificity (in terms of non-dropout accuracy) is given by tn/(fp+tn). The classifier with a larger 'area under the ROC curve' (AUROC) proves a better performance than that with a smaller AUROC.

5.3 Cross validation

A 10-fold cross validation (10-fold CV) procedure is conducted in the experiments for each week. During the 10-fold CV procedure, each data set is first randomly divided into 10 subsets containing an equal number of samples, where 9 subsets are used to train the classifier and the remaining is used for testing. The classification result (e.g., *F-score*) is then obtained on each testing data set of the cross validation. Afterwards, results obtained from the 10-fold CV are averaged. Note that data from the 227 participants without missing values and outliers are used to select features with GA prior to CV. And all the 950 participants are included in the 10-fold CV due the limited number of dropout samples per week.

5.4 Statistical comparison

As shown in Table 1, we considered 6 cases using different features and/or methods for comparison. The significance of difference between *F-score* values obtained in different cases can be examined via an (1-sided) paired Wilcoxon signed-rank test (here at $p < 0.001$).

6 Results

As defined in Section 1.4, participants in the program can be classified into $dropout$ and $non\text{-}dropout$. Figure 5 illustrates the distribution of the dropout/non-dropout rate (%) over the 11 data sets from week 1 to 11, where for each week, it is computed as the ratio of the number of dropouts in that week to the total number of all non-dropouts (amount to 559 out of the 950 participants). So in our data sets, dropouts only account for 5.9 % per week on average. Many participants (47) dropped out in the first week, and 16 participants did not really start the program after the AW or did not enter the AW at all. The reasons for these early dropouts are unknown.

Table 2 presents the discriminative power of the features in separating $dropout$ and $non\text{-}dropout$ classes of every future week. Since there are a large number of features, only the top 3 features (ranked by D_H) are given in the table. On average, the features daily wearing hours, percentage of wearing days and average PAL score during the past IWs performed the best. Note that it does not mean that the other relatively less-discriminative features are useless, they would still contribute additional information to predicting future dropouts when combined in a certain way.

Table 3 summarizes the classification performance in different cases (see Table 1) when using a NB and a KNN classifier. When comparing the results of case 1 with case 2, we notice an improvement of ~0.02 in *F-score* after combining the ARIMA features for both classifiers. Generating and selecting features through GA from the traditional

Table 1 Cases with different features and/or methods for comparison

	Trad. features	ARIMA features	GA	PCA	WCSL
Case 1	×				
Case 2	×	×			
Case 3	×	×		×	
Case 4	×	×	×		
Case 5	×	×	×	×	
Case 6	×	×	×	×	×

Note: the traditional feature set consist of *personal*, *assessment*, *dynamic* and *global* features. For case 3, 5 and 6, four dimensions of feature vectors after PCA were selected.

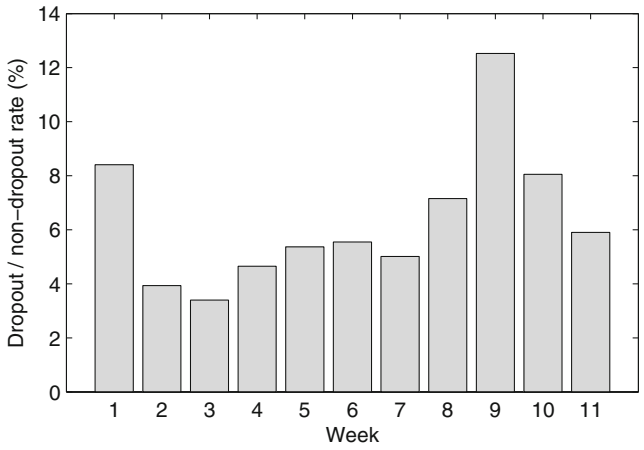

Fig. 5 Dropout/non-dropout rate (%) versus week in the program

and ARIMA feature set resulted in a further increase in *F-score* of ∼0.02 (see the differences between case 4 and case 2). Note that a total of 40 features were experimentally obtained after GA. Besides, slight increases of *F-score* can be observed in the table after using a WCSL method (by comparing case 6 with case 5). This method also helps balance the results of sensitivity and specificity by assigning different costs of misclassification (adjusted by the weight ω) to the unbalanced classes. The *F-score* scored highest when $\omega = 0.45$ for the two classifiers.

To reduce the dimensionality in feature space via PCA, we computed the average *F-score* using 10-fold CV with different numbers of chosen "components" (i.e., dimensions), with the selection priority on the ranking of their corresponding eigenvalues listed in a descending order. The results on the training sets during 10-fold CV indicate that both classifiers using the first four components after PCA already provided an optimal performance. Thus there was definitely redundancy in the features and PCA can largely reduce the complexity of calculation. The performance obtained using more components did not improve further, and instead dropped to some extent. The reason might be that the remaining components barely contributed to the classification and the useful information carried was limited compared to the increase of noise level. Compared with the results obtained without PCA (case 2 and 4), slight improvements are observed when using PCA (case 3 and 5). This means that PCA also served to increase the overall discriminative power of the features to a certain degree and thus lead to a further enhancement of the classification performance. Note that all the features were normalized to have zero mean and unit variance prior to PCA, which aimed at ensuring that they were scaled comparatively.

Compared with the original results obtained in case 1, the use of GA, PCA and WCSL (case 6) achieved a significantly improved *F-score* of 0.21 ± 0.09 (at a sensitivity of 63.0 % and a specificity of 70.9 %) with a NB classifier and of 0.22 ± 0.08 with a KNN classifier (at a sensitivity of 66.4 % and a specificity 74.1 %). We note that the sensitivity and specificity (case 6) are both increased when using NB and become more balanced when using KNN.

In addition, the KNN classifier performs slightly better than NB after applying the proposed WCSL method. This

Table 2 Discriminative power of the top-ranked features as measured by the Hellinger distance metric (D_H)

Week	Features (Top 3)	D_H
	Average logarithm PAL	0.38
1†	Average PAL score on Saturday	0.36
	Moderate-intensity PAL minutes*	0.34
	Wearing hours of past IWs*	0.52
2	Wearing hours of AW*	0.49
	PAL seasonality of past IWs	0.37
	Wearing hours of past IWs*	0.54
3	Average PAL score of past IW	0.47
	PAL seasonality of past IWs	0.44
	Wearing hours of past IWs*	0.41
4	Percentage of wearing days of past IWs	0.39
	Age	0.38
	High-intensity PAL minutes before week 4*	0.44
5	Percentage of wearing days of past IWs	0.41
	High-intensity PAL minutes of past IWs*	0.37
	Wearing hours of past IWs*	0.41
6	Average PAL score before week 3	0.40
	Percentage of wearing days before week 5	0.39
	Percentage of wearing days before week 5	0.42
7	Moderate-intensity PAL minutes of past IWs*	0.41
	Ratio of activity to target of past IWs*	0.40
	Moderate-intensity PAL minutes of past IWs*	0.40
8	Moderate-intensity PAL minutes before week 3*	0.39
	Starting month	0.35
	Starting month	0.38
9	Maximal PAL day	0.27
	LED activation times before week 7*	0.26
	Ratio of activity to target of past IWs*	0.38
10	Average PAL score of past IWs	0.37
	Ratio of activity to target before week 8*	0.36
	Percentage of wearing days before week 9	0.29
11	Percentage of wearing days before week 8	0.29
	Wearing hours before week 5*	0.27
	Wearing hours of past IWs	0.34
Average‡	Percentage of wearing days of past IWs	0.30
	Average PAL of past IWs	0.28

† The features were extracted from the data in the AW

‡ Average over weeks

* The feature was computed as the average of daily measures

Table 3 Summary of classification performances using NB and KNN classifiers for different cases

	Classifier	Sensitivity†	Specificity‡	*F-score*
Case 1	NB	59.4 ± 9.2 %	61.6 ± 10.9 %	0.15 ± 0.07
	KNN	18.8 ± 9.7 %	90.7 ± 5.3 %	0.13 ± 0.06
Case 2	NB	61.4 ± 9.9 %	63.2 ± 10.6 %	0.17 ± 0.08
	KNN	20.7 ± 10.1 %	91.0 ± 5.9 %	0.15 ± 0.07
Case 3	NB	62.2 ± 10.2 %	63.8 ± 10.5 %	0.18 ± 0.08
	KNN	21.9 ± 10.6 %	91.4 ± 6.1 %	0.16 ± 0.08
Case 4	NB	60.3 ± 13.8 %	70.5 ± 11.0 %	0.19 ± 0.09
	KNN	24.2 ± 12.7 %	89.3 ± 5.2 %	0.17 ± 0.09
Case 5	NB	61.1 ± 13.5 %	72.1 ± 10.8 %	0.20 ± 0.09
	KNN	26.6 ± 12.3 %	90.5 ± 4.7 %	0.19 ± 0.08
Case 6	NB	63.0 ± 13.9 %	70.9 ± 11.4 %	0.21 ± 0.09
	KNN	66.4 ± 13.8 %	74.1 ± 7.3 %	0.22 ± 0.08

† Dropout accuracy
‡ Non-dropout accuracy

can also be easily recognized in the ROC space in Fig. 6, where the optimal results are marked. It should be pointed out that the choice of k in the KNN classifier is important since it may affect the classification performance. The k value of 21 was chosen to optimize the average *F-score* over all the training sets during 10-fold CV. The use of the WCSL method solved the unbalanced data set problem and then provided an improved result so that the achieved accuracies of the two classes became balanced. In other words, it can largely improve the sensitivity by sacrificing some specificity. Moreover, as indicated in Table 4 (case 6), the WCSL method can help KNN more, which originally performed worse in classifying unbalanced data sets. Both of the classifiers perform well above random guess (with an accuracy of 50 %). Table 4 reports the details of the improved classification achieved using a KNN classifier (case 6). It does not make a lot of sense to compare the *F-score* over the 11 weeks because of their different sample sizes and prior probabilities of dropping out in different weeks.

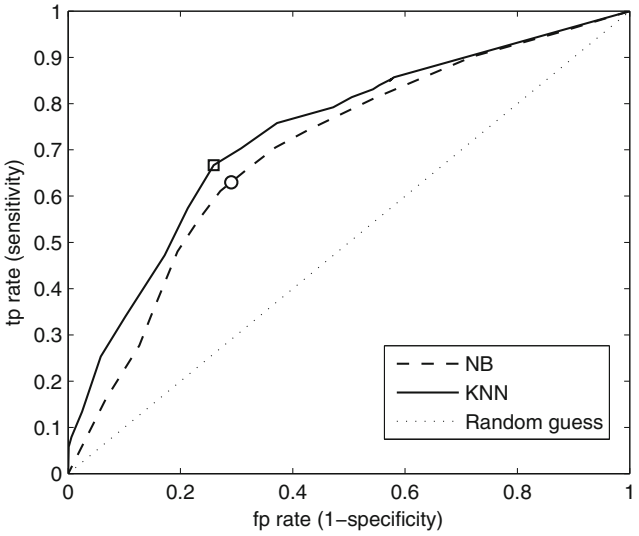

Fig. 6 Performance comparison of the two classifiers using the combination of the traditional and the ARIMA features (with GA and PCA) in the ROC space after applying the WCSL method (case 6). The square and circle markers represent the optimal points for the KNN and NB classifiers, respectively, as measured by *F-score*. The dotted line represents results of 'random guess', indicating that the two classes have no discrimination

7 Discussion

7.1 Classification

The dropout and non-dropout prediction accuracies by using a NB classifier are more balanced than those obtained with a KNN classifier when comparing the results without using WCSL in Table 3. This is because, for NB, the density function of the minority class can still be estimated continuously when the number of samples in

Table 4 Classification results per week with a KNN classifier after using GA, PCA, and WCSL based on 10-Fold CV

Week	Sensitivity	Specificity	*F-score*
1	63.8 %	81.9 %	0.32
2	95.5 %	68.0 %	0.19
3	84.2 %	64.0 %	0.13
4	46.2 %	78.5 %	0.14
5	53.3 %	79.6 %	0.20
6	67.7 %	71.7 %	0.20
7	57.1 %	80.1 %	0.20
8	65.0 %	73.0 %	0.24
9	65.7 %	83.8 %	0.43
10	71.1 %	62.1 %	0.22
11	60.6 %	72.3 %	0.19
Average	66.4 ± 13.8 %	74.1 ± 7.3 %	0.22 ± 0.08

this class is not too low, though the density function estimation of the minority class may not be as accurate as that of the majority class. It means that the density function estimation of NB is robust against an unbalanced data set if the training sample sizes of both classes are large enough. However, for KNN, the densities of the two classes in the area having the nearest k neighbors were seriously skewed due to the class unbalance. In fact, it has been reported that KNN is sensitive to an unbalanced data set [60].

Regarding the classifiers, some other algorithms have been studied on predicting future events based on historical data. For instance, Pentland et al. [27] proposed a Markov dynamic model (MDM) with expectation-maximization (EM) to predict driver actions, which outperforms nearest-neighbor, Bayesian and hidden Markov models. To predict the loss of customers for banks, an improved balanced random forest (IBRF) algorithm achieved a significantly higher prediction accuracy compared with the artificial neural network and decision tree [62]. The 'cost-sensitive version' of a support vector method (SVM) that optimizes *F-score* offered an improved performance, in particular for text classification with highly unbalanced classes [63]. However, these conclusions might not hold in our study. It is hence suggested to further investigate and compare the classification algorithms based on our database, where the two classes are unevenly distributed.

In general, the prediction results are not as good as we expected, in which the *F-score* is only 0.22. There might be a 'ceiling' of the accuracy of classifying dropouts and non-dropouts. The dropout behaviors of some participants are hardly predictable when they behaved consistently before they actually dropped out. For instance, they decided to give up using the monitors occasionally or they lose their devices rather than they intended to drop out. Even though, the performance still has some opportunities of being improved by employing different prediction algorithms, as discussed before. On the other hand, the PAL data along with corresponding behavioral information derived from uploaded data might be affected by problems when using the device. For instance, a flat battery and lying, sitting or standing without any activity might lead to inaccurate estimates of the feature daily wearing hours, yielding errors in predicting dropouts. These problems need further investigation.

7.2 Definition of dropouts

The accuracies score fairly consistently over the 11 weeks ($\geq$65 %), with the exceptions of weeks 1, 4, 5, 7, and 11 for the dropout prediction and of weeks 3 and 10 for the non-dropout prediction (see Table 4). A possible reason of low dropout accuracy in week 11 might be in the definition of a dropout. The dropouts in week 11 are close to the decision boundary which separates dropouts from non-dropouts. A participant who stopped registering PAL data would not be considered a dropout at this stage of the program if they stop registering activity only one day later. Therefore, such an arbitrary decision boundary might be algorithmically difficult to distinguish between participants close to it. The reasons for the other weeks with low classification accuracies are unclear. These might be because the dropout was not very well-defined, as long as there is no formal definition of it and participants did not say that they had quit the program prematurely. Hence, a more well-defined criteria for dropout is required.

7.3 Variances between weeks

As presented in Table 1, the most discriminative features vary over weeks. With respect to time, the top-ranked features are those often obtained based on all past IWs or the week near the target week. And the *assessment* features are more helpful to the dropout prediction in the beginning of the program (e.g., week 1 and 2). These indicate that their discriminative powers might be decreasing over time. The feature starting month is highly ranked for week 8 and 9, which might be because many participants started the program from November and possibly dropped out and went for their Christmas holidays in those two weeks. So for these two weeks, this feature can well distinguish between them and the other non-dropout participants who started the program earlier.

In addition, some weeks show considerable differences between dropout and non-dropout (prediction) accuracies, such as week 2, 4, 5, and 7 (see Table 4). The WCSL method was applied to all the 11 data sets with the same weight ω but the optimal results for different weeks could differ when choosing different ω. However, using different weights for the classifications of different weeks would lead to overfitting to the data sets. Furthermore, it was assumed that the distribution of dropout rates over the whole program is uniform (with a same prior probability of dropout versus week), but this might not hold as long as the number of weeks is not sufficient in the program to examine this assumption. Accepting it might result in incorrectly optimizing the accuracies of predicting dropouts for different weeks. Therefore, this assumption should be further examined based on the data from more weeks.

The results also indicate a large variance of *F-score* between weeks (from 0.13 to 0.43) with a relatively high standard deviation of 0.08. This implies that the PAL scores vary across participants and the selected features and classifiers might not be optimized for a specific participant or week. To eliminate this, specialized features and classifiers for different participants or weeks are suggested.

7.4 Limitations of our study or methods

The results provided in Section 6 show that the classification performance can be improved through combining the ARIMA features with the traditional features and incorporating GA, PCA, and WCSL. However, there are some limitations or disadvantages.

- During the PAL data modeling process, only a small subset of the database (227 out of 950 participants) was analyzed in this study. This is because, as described in Section 3, including the data from the other participants (with missing values and outliers) for PAL data fitting might affect the adequacy of our ARIMA-based models. Therefore, the models and the extracted ARIMA features might not reflect the common properties of the population. It is suggested to address a larger-sized data set in future work.
- The discriminative powers of the personal characteristic features (e.g., age and gender) presented in Table 2 are not as high as we expected when compared with the other top-ranked features, which seems not consistent to previous findings [64]. This might be because in this study we only collected data from the company employees (see Section 2.3) who cannot represent the whole population. The use of a larger-sized data set including participants with other professions should be further considered.
- The errors of predicting future PAL data might exist in the features extracted from the predicted PAL data, which would therefore introduce noise to the dropout/non-dropout predictor. Such 'error propagation' merits further investigation.
- The automated C-ARIMA process would result in a higher computation load. Similarly, genetic programming is also computationally intensive, especially when the number of participants and features in the database increases. Fortunately, the computational costs may remain manageable as long as only a single run of the ARIMA and GA processes on the training data per week is required for classification.
- Using GA that creates combinations of existing features not only results in an unclear interpretation of the contributions of the original features on the final classification performance, but also complicates the matter of providing interventions to the possible dropout participants by a human coach.
- The *F-score* might not be the most suitable measure to optimize the classification performance. It mainly depends on the requirement or preference of sensitivity (or dropout accuracy) and specificity (or non-dropout accuracy) in a real application. For instance, to prevent as many participants from dropping out of the program as possible, interventions should be delivered to a lot of participants even though some of them do not intend to drop out. In this situation, the dropout accuracy is required to be higher even if the non-dropout accuracy will become lower. To achieve this, the recall in the *F-score* equation should be higher weighted than the precision. However, a too low non-dropout accuracy will drive the human coach to wrongly deliver interventions to the participants who are not at high risk of dropping out. Hence, it merits further investigating the selection of evaluation criterion from a practical perspective.
- Since the participants were not asked to provide information about dropping out of the physical activity intervention program, the reasons of dropouts were unknown in this study. This would yield difficulties in understanding the features and the classifiers for predicting dropouts and then in providing them effective interventions. Thus, this needs to be studied in the future.

8 Conclusion

In this study, participants were classified every week as either future dropout or non-dropout, corresponding to their risk of early dropout of a 12-week physical activity intervention program based on the PAL data acquired from a daily wearable sensor – activity monitor. A KNN classifier achieved prediction accuracies of ∼66 % and ∼74 % for dropout and non-dropout, respectively. The addition of the ARIMA features to the feature set and the use of GA yielded a clear improvement in classification accuracy. The use of a WCSL method in the classifiers to fight against the unbalanced data set improved the classification performance further. A more accurate prediction of a likely dropout case provides a coach with insight into which of the participants is at the highest risk for future dropping out. Based on this insight, the coach can direct his/her efforts to those individuals that have the strongest need for coaching support, allowing the coach to intervene timely and effectively to motivate these participants to stay in the program, towards a tailored physical activity health intervention. However, these interventions can only be delivered to the dropouts who are correctly predicted. The number of dropout participants who will be successfully prevented from dropping out is unknown based on the results in this study, it usually depends on the interventions, which merits further investigation.

Conflict of interests The authors declare that they have no conflict of interest.

Acknowledgements The authors would like to thank three anonymous reviewers, and Dr. A. Bonomi, Dr. R. Haakma, and Dr. S. Jelfs from Philips Research Laboratories for their insightful comments.

References

1. Penedo FJ, Dahn JR. Exercise and well-being: review of mental and physical health benefits associated with physical activity. Curr Opin Psychiatry 2005;18(2):189–93.
2. Pate RR, Pratt M, Blair SN, Haskell WL, Macera CA, Bouchard C, Buchner D, Ettinger W, Heath GW, King AC. Physical activity and public health: A recommendation from the centers for disease control and prevention and the American college of sports medicine. J Am Med Assoc 1995;273(5):402–7.
3. Fox KR. The influence of physical activity on mental well-being. Public Health Nutr 1999;2(3a):411–8.
4. Driver HS, Taylor SR. Exercise and sleep. Sleep Med Rev 2000;4(4):387–402.
5. Ware LJ, Hurling R, Bataveljic O, Fairley BW, Hurst TL, Murray P, Rennie KL, Tomkins CE, Finn A, Cobain MR, Pearson DA, Foreyt JP. Rates and determinants of uptake and use of an internet physical and weight management program in office and manufacturing work sites in England: cohort study. J Med Internet Res 2008;10(4):e56.
6. Goris AHC, Holmes R. The effect of a lifestyle activity intervention program on improving physical activity behavior of employees. In Proc. 3rd Int Conf Persuasive Technol (PERSUASIVE), Oulu, Finland. 2008.
7. Lacroix J, Saini P, Holmes R. The relationship between goal difficulty and performance in the context of a physical activity intervention program. In Proc 10th Conf MobileHCI, Amsterdam, Netherlands. 2008;415–8.
8. Plasqui G, Joosen AM, Kester AD, Goris AHC, Westerterp KR. Measuring free-living energy expenditure and physical activity with triaxial accelerometry. Obes Res 2005;13(8):1363–9.
9. Hurling R, Catt M, Boni MD, Fairley BW, Hurst T, Murray P, Richardson A, Sodhi JS. Using internet and mobile phone technology to deliver an automated physical activity program. Randomized controlled trial. J Med Internet Res 2008;9(2):e7.
10. Clarkson BP. Life patterns: structure from wearable sensor. Ph.D Thesis: MIT Media Lab; 2002.
11. Long X, Yin B, Aarts RM. Single-accelerometer-based daily physical activity classification. In Proc 31st Ann Int Conf IEEE EMBS, Minneapolis, MN, 2009;6107–10.
12. Bouten C, Westerterp K, Verduin M, Janssen J. Assessment of energy expenditure for physical activity using a triaxial accelerometer. Med Sci Sports Exerc 2003;26(12): 1516–23.
13. Bonomi AG, Plasqui G, Goris AHC, Westerterp KR. Estimation of free-living energy expenditure using a novel activity monitor designed to minimize obtrusiveness. Obes 2010;18 (9):1845–51.
14. Leenders NY, Sherman WM, Nagaraja HN, Kien CL. Evaluation of methods to assess physical activity in free-living conditions. Med Sci Sports Exerc 2001;33(7):1233–40.
15. Marcus BH, Lewis BA, Williams DM, Dunsiger S, Jakicic JM, Whiteley JA, Albrecht AE, Napolitano MA, Bock BC, Tate DF, Sciamanna CN, Parisi AF. A comparison of internet and print-based physical activity interventions. Arch Intern Med 2007;167(9):944–9.
16. van den Berg MH, Schoones JW, Vliet Vlieland TP. Internet-based physical activity interventions: a systematic review of the literature. J Med Internet Res 2007;9(3):e26.
17. Adams J White M. Are activity promotion interventions based on the transtheoretical model effective? A critical review. Br J Sports Med 2003;37(2):106–14.
18. Marcus BH, Bock BC, Pinto BM, Forsyth LH, Robeerts MB, Traficante RM. Efficacy of an individualized, motivationally-tailored physical activity intervention. Ann Behav Med 1998;20(3):174–80.
19. Segerstahl K, Oinas-Kukkonen H. Distributed user experience in persuasive technology environments. In Proc. 2nd Int Conf Persuasive Technol (PERSUASIVE), Palo Alto, CA. 2007.
20. Xu W, Zhang M, Sawchuk AA, Sarrafzadeh M. Robust Human activity and sensor location co-recognition via sparse signal representation. IEEE Trans Biomed Eng 2012;59(11):3169–76.
21. Khan AM, Lee YK, Kim TS. A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. IEEE Trans Inf Technol Biomed 2010;14(5):1166–72.
22. Zhu C, Sheng W. Realtime recognition of complex human daily activities using human motion and location data. IEEE Trans Biomed Eng 2012;59(9):2422–30.
23. Mazilu S, Hardegger M, Zhu Z, Roggen D. Online detection of freezing of gait with smartphones and machine learning techniques. In Proc 6th Int Conf PervasiveHealth 2012:123–30.
24. Lustrek M, Kaluza B. Fall detection and activity recognition with machine learning. Informatica 2009;33:205–12.
25. Jin A, Yin B, Morren G, Duric H, Aarts RM. Performance evaluation of a tri-axial accelerometry-based respiration monitoring for ambient assisted living. In Proc. 31st Ann Int Conf IEEE EMBS, Minneapolis, MN, 2009;5677–80.
26. Nijsen TME, Aarts RM, Cluitmans PJM, Griep PAM. Time-frequency analysis of accelerometry data for detection of myoclonic seizures. IEEE Trans Inf Technol Biomed 2010;14(5):1197–203.
27. Pentland A, Liu A. Modeling and prediction of human behavior. Neur Comp 1999;11:229–42.
28. Yang M, Wong SCP, Coid J. The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. Psych Bullet 2010;136(5):740–67.
29. Cheng S, Tom K, Thomas L, Pecht M. Wireless sensor system for prognostics and health management. IEEE Sensors J 2010;10(4):856–62.
30. Long X, Pauws S, Pijl M, Lacroix J, Goris AHC, Aarts RM. Predicting daily physical activity in a lifestyle intervention program. In: Gottfried B, Aghajan H, editors. Behaviour Monitoring and Interpretation - Well-Being. The Netherlands: IOS Press; 2011, pp. 131–46.
31. Si XS, Hu CH, Yang JB, Zhou ZJ. A new prediction model based on belief rule base for system's behavior prediction. IEEE Trans Fuzzy Syst 2011;19(4):636–51.
32. Schwabacher M, Goebel K. A survey of artificial intelligence for prognostics. In Proc AAAI Fall Sym, Arlington, VA. 2007;107–14.
33. Niu G, Yang BS. Dempster-Shafer regression for multi-step-ahead time-series prediction towards data-driven machinery prognosis. Mech Syst Sig Process 2009;23:740–51.
34. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intell Data Anal 2002;6(5):429–49.
35. Ling CX, Li C. Data mining for direct marketing: problems and solutions. In Proc 4th ACM SIGKDD Int Conf Knowl Disc Data Min 1998:73–9.
36. Zhang J, Mani I. kNN approach to unbalanced data distributions: a case study involving information extraction. In Proc ICML Workshop Learning from Imbalanced Datasets. Washington, DC. 2003.

37. Bradley AP. The use of area under the ROC curve in the evaluation of machine learning algorithms. Patt Recogn 1997;30(7):1145–59.
38. Duda R, Hart P, Stork D. Pattern Classification, 2nd ed: Wiley; 2001.
39. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory 1967;13(1):21–7.
40. Hand DJ, Yu K. Idiot's Bayes – not so stupid after all. Int Stat Review 2001;69(3):385–98.
41. Bhatia N. Vandana. Survey of nearest neighbor techniques. Int J Comput Sci Inf Secur 2010;8(2):302–5.
42. Weiss GM, Learning with rare cases and small disjuncts. In Proc 12th Int Conf Mach Learn LakeTahoe CA. 1995;558–65.
43. Kubat M, Matwin S. Addressing the curse of imbalanced datasets: One-sided sampling. In Proc 14th Int Conf Mach Learn 1997:179–86.
44. Pazzani M, Merz C, Murphy P, Ali K, Hume T, Brunk C. Reducing Misclassification Costs. In Proc 11th Int Conf Mach Learn 1994:217–25.
45. Yang J, Honavar V. Feature subset selection using a genetic algorithm. IEEE Trans Intell Syst App 1998;13(2):44–9.
46. Bonomi AG, Plasqui G, Goris AHC, Westerterp KR. Improving assessment of daily energy expenditure by identifying types of physical activity with a single accelerometer. J Appl Physiol 2009;107(3):655–61.
47. Shetty PS, Henry CJ, Black AE, Prentice AM. Energy requirements of adults: an update on basal metabolic rates (BMRs) and physical activity levels (PALs). Eur J Clin Nutr 1996;50(1):S1—23.
48. World Health Organization. Human energy requirements. Rome: Report of a Joint FAO/WHO/UNI Expert Consultation; 2011.
49. Clements MP, Hendry DF. Forecasting economic time series. Cambridge: Cambridge University Press; 1998.
50. Helfenstein U. Box-Jenkins modelling in medical research. Stat Meth Med Res 1996;5(1):3–22.
51. Peng JY, Aston JAD. The state space models toolbox for MATLAB. J Stat soft 2011;41(6):1–26.
52. Box GEP, Jenkins GM. Time series analysis forecasting and control. San Francisco: Holden-Day; 1976.
53. Bowerman BL, O'Connell RT. Forecasting and time series: an applied approach. Belmont: Duxbury Press; 1993.
54. Harvey AC. Forecasting, structural time series models and the Kalman filter. Cambridge: Cambridge University Press; 1989.
55. Schwarz G. Estimating the dimension of a model. Ann Stat 1978;6(2):416–64.
56. Koza J. Genetic programming: on the programming of computers by means of natural selection: MIT Press; 1992.
57. Abdi H, Williams LJ. Principal component analysis. WIREs: Comp Stat 2010;2(4):433–59.
58. Parzen E. On estimation of a probability density function and mode. Ann Math Stat 1962;333(3):1065–76.
59. Hellinger E. Neue Begründung der Theorie quadratischer formen von unendlichvielen veränderlichen. J für die Reine und Angew Math 1909;136:210–71.
60. Provost F, Fawcett T. Analysis and visualization of classifier performance: comparison under imprecise class and cost distribution. In Proc 3rd ACM SIGKDD 1997:43–48.
61. van Rijsbergen CJ. Information retrieval. London: Butterworths; 1979.
62. Xie Y, Li X, Ngai EWT, Ying W. Customer churn prediction using improved balanced random forests. Exp Syst Appl 2009;36(3):5445–9.
63. Joachims T. A support vector method for multivariate performance measures. In Proc 22nd Int Conf Mach Learn 2005:377–84.
64. King AC, Rejeski WJ, Buchner DM. Physical activity interventions targeting older adults: a critical review and recommendations. Am J Prev Med 1998;15(4):316–33.