



# Unsupervised statistical concept drift detection for behaviour abnormality detection

Björn Friedrich<sup>1</sup> · Taishi Sawabe<sup>2</sup> · Andreas Hein<sup>1</sup>

Accepted: 9 April 2022 / Published online: 10 May 2022  
© The Author(s) 2022

## Abstract

Abnormal behaviour can be an indicator for a medical condition in older adults. Our novel unsupervised statistical concept drift detection approach uses variational autoencoders for estimating the parameters for a statistical hypothesis test for abnormal days. As feature, the Kullback–Leibler divergence of activity probability maps derived from power and motion sensors were used. We showed the general feasibility (min.  $F_1$ -Score of 91 %) on an artificial dataset of four concept drift types. Then we applied our new method to our real-world dataset collected from the homes of 20 (pre-)frail older adults (avg. age 84.75 y). Our method was able to find abnormal days when a participant suffered from severe medical condition.

**Keywords** Healthcare · Behaviour analysis · Unsupervised concept drift detection · Older adults · Ubiquitous computing

## 1 Introduction

Behaviour monitoring for healthcare purposes is an important task and a change in behaviour may indicate a medical condition. This includes physical diseases as well as cognitive diseases like dementia [1–4]. To get a holistic view of the behaviour, unobtrusive privacy preserving sensors can be installed in domestic environments. Human behaviour is very volatile and difficult to analyse, but machine learning approaches showed some promising results in this domain [5]. One disadvantage is that a vast amount of data is needed for training, thus in some cases, collecting large datasets is expensive or even impossible. Moreover, labeling the data is time consuming and labels can be unreliable.

Unsupervised machine learning algorithms are not relying on labels but need more training data. Our real-world dataset, collected during the 10-months observational OTAGO study, suffers from both limitations. Smart home sensors were installed in the domestic environments of 20 (pre-)frail older adults aged 85.75 y (SD: 5.19 y) and diaries were written for each participant. The dataset is both small and has unreliable labels, that is why we noticed that the most promising approach was to use unsupervised machine learning. Our goal was to detect changing or drifting behaviour and abnormal behaviour. Concept drift detection is one approach for reaching our goal. considering the limitations of our dataset the unsupervised concept drift detection seemed to be a reasonable approach. We extended the common approach by a novel method to deal with small datasets. We used the Central Limit Theorem (CLT) and the generalisation abilities of Variational Autoencoders (VAE) to get an accurate estimation of the underlying random distribution of our feature set. The feature set contains the Kullback–Leibler Divergence ( $D_{KL}$ ) of daily activity probability maps and a 7 d baseline activity probability map [6]. If a new day was normal the baseline was updated with the new day. Our new method can be applied to all features that are satisfying the CLT assumptions. Our aim was to detect abnormal behaviour as indication for medical condition in older adults. Before applying our new method to our real-world dataset we validated it in two steps on an artificial datasets. The first step was to validate whether the VAE can approximate the underlying distribution more

---

✉ Björn Friedrich  
bjoern.friedrich@uni-oldenburg.de

Taishi Sawabe  
t.sawabe@is.naist.jp

Andreas Hein  
andreas.hein@uni-oldenburg.de

<sup>1</sup> Department of Health Services Research, Carl von Ossietzky University, Ammerländer Heerstraße 114-118, Oldenburg, 26129, Lower-Saxony, Germany

<sup>2</sup> Interactive Media Design Laboratory, Nara Institute of Science and Technology, 8916-5, Takayama-cho, Ikoma-shi, 630-0192, Nara-ken, Japan

accurately than common approach using the sample mean and the sample variance. In the second step we show that our approach can detect the four basic concept drift detection types. After our method passed both validation steps we applied it to our real-world dataset to detect abnormal behaviour of older adults. In the next section we are giving an overview of unsupervised concept drift detection methods and machine learning approaches for behaviour anomaly detection. In Section 3 we establish the theoretical foundations of our approach and we describe the data acquisition process, the preprocessing steps, feature engineering, and our validation process in detail. In the results Section 4 we present our results and we discuss them in the following section. In the last Section 6 we draw conclusions and provide an outlook for future work.

## 2 State of the art

In the field of machine learning several approaches using different algorithms have been introduced. Common approaches with good results for abnormality detection and activity recognition in smart homes were using Long Short-Term Memory networks, which take the time domain into account [7–9]. However, these supervised approaches are dependent on labeled data. In contrast, the approaches developed in [10–12] were probability density based. A statistical hypothesis test was used to determine if a new sample or set of samples was drawn from the same distribution. The authors in [13] combined a statistical significance test with the *k*-Nearest-Neighbour (kNN) algorithm. The kNN algorithm was used to partition the samples in the sample space. The significance test was used to check if a drift was present. A Student-Teacher approach for unsupervised concept drift detection was introduced in [14]. The student model was trained to mimic the loss of the teacher model. When new unlabeled samples arrived the discrepancy between the student and teacher models were used as surrogate signal for the concept drift detection. A common unsupervised concept drift detection method was used for the final detection step. With time series data in mind [15] developed an algorithm which was a composition of three different steps, dimensionality reduction, random partitioning, and anomaly detection. The anomalies were detected based on frequency statistics and the proportion of current anomalies and the average of previously occurred anomalies. The average is updated over time and leads to an adoption of the algorithm. The afore mentioned research was not focused on medical and healthcare applications. In healthcare, particularly for the task of abnormal behaviour detection, unsupervised concept drift detection had been used. Using motion sensor data, a graph of the flat, and a transition matrix of an unsupervised behaviour monitoring

system was developed in [16]. The features were the probability of room-to-room transitions and detention time in a room. If a person made a transition or stayed in a room with a low probability an anomaly was detected. The probabilities were updated over time and able to adapt to the person's behaviour. The method introduced in [17] focused on the drift in performing activities of daily living. A similarity matrix between all time windows consisting of a sequence of raw sensor events were computed. The similarity was defined as the intersection of two histograms derived from the time windows. Afterwards the results were clustered using Markov Clustering. For detecting if a new time window is a drift they used the Silhouette Index in combination with a threshold. In [18] the aim was to assess the risk of abnormalities, that is why rule-mining was used to identify the causes of abnormalities. The resultant rules were used in combination with a Markov Logic network to detect the risk of abnormalities. Temporal information for abnormality detection were used in [19]. Based on sensor events and activities logs the probabilities of the temporal relation of activities were identified, e.g. the probability of one activity happens after a certain other activity. If a combination with a low probability occurred a drift was detected. The approach introduced in [20] was based on an unsupervised Long-Short-Term-Memory (LSTM) autoencoder. The encoder was trained on a baseline dataset containing motion sensor data, then the trained autoencoder was used to identify abnormalities by evaluating the deviation from the learned regular model. The authors in [21] tried to detect abnormal days by comparing the activities performed on the day to the previous days. In the first step they used a machine learning approach to classify the activities based on smart home sensor data and in the second step they computed the boundary between normal and abnormal days by expanding the set of normal days. When the expansion stops the model was used for further classification. A more general review in [22] found that unsupervised deep learning methods outperform the (semi-) supervised methods in the task of abnormality detection in human behaviour.

We contribute an approach using a VAE as basis for a statistical test for concept drift and accurately estimating the parameters for a probability distribution for small datasets to the field of machine learning. We contribute to the healthcare domain by applying our method to unstructured smart home data of 20 (pre-)frail older adults for anomaly detection.

## 3 Materials and methods

### 3.1 Data acquisition

The data of 20 participants (17 female, 3 male) was collected during the observational OTAGO study conducted

by the University of Oldenburg in 2014 and 2015 over a period of 10-months [23]. Due to dropouts the average participation time became 36.5 weeks. The average age of the participants was 84.75 y (SD: 5.19 y) and they were pre-frail or frail by the definition of Fried [24, 25]. The participants lived alone in their own flats, but there might be a possibility that they were assisted by a care giver. Every four weeks the staff from the University of Oldenburg visited the participants for supervising geriatric assessments, completing questionnaires, and a diary. Holidays and time conflicts led to an average of 31.3 d (SD: 5.3 d) between two the visits. The following assessments for assessing the mobility and physical function were performed Tinetti [26], Short Physical Performance Battery (SPPB) [27], Timed Up & Go (TUG) [28], and Hand Grip Strength (HGS) [29]. For assessing the independence, quality of life, and nutritional status the questionnaires Instrumental Activities of Daily Living (IADL) [30], EQ-5D-5L [31], and Mini-Nutritional Assessment (MNA) [32] were completed. The EQ-5D-5L and the MNA were only completed twice, at the beginning and the end of the study. Tables 1 and 2 show selected characteristics of the cohort at the beginning of the study (T0) and the end of the study (T10). In addition to the regular visits and assessments, a variety of sensors were installed in the flats of the participants.

The sensor set contained home automation and power consumption sensors. Additionally, two wearable sensors, Shimmer3r (Inertial Measurement Unit) [33] and Columbus V990 (GPS) [34], were given to the participants. A four key switch was installed next to the main entrance door and the participants were asked to turn it, when other people were entering or leaving the flat. The power consumption sensors were attached between the appliance and the power socket and measured all energy consumption of the appliance in W. The home automation sensors were comprised of passive infrared motion (PIR) sensors, contact sensors, and concussion sensors. The motion sensors were installed in the rooms to detect motion in a certain area and they had a scan-dead time of 8 s. Some sensors were installed to measure specific walk paths inside the flats, e.g. the way from the bedroom to the kitchen. One motion sensor was placed right over the lavatory flush to detect toilet use. The contact sensors were installed at all entrance doors of the flats and at the fridge. All home automation sensors had a wireless connection to a base station to transmit their data. The sensor setup was dependent on the topology and the

wishes of the participants. A flat of a participant of the OTAGO study is shown in Fig. 1.

### 3.2 Data preprocessing

Data preprocessing is an important step in machine learning application. That holds especially for medical data, where very often values are missing and samples are incomplete. Extensive preprocessing using imputation methods showed good results [35]. However, the concussion sensor was identified as unreliable and faulty in a way that imputation methods were not applicable. Therefore, we excluded it from the further analysis. One participant reported a fallen motion sensor and the sensor was fixed after. We did not exclude the data, because we were also interested in how the results were affected. The main preprocessing step was filtering and resampling the power sensor consumption data. The power consumption sensors were measuring standby energy consumption as well. Since the standby state did not contain any information about activity of a participant, standby times were filtered. We assumed that the appliances were in standby most of the time. To find the standby power consumption of each appliance, we used a histogram to find the dominating W consumption and considered all values smaller or equal as standby power consumption. The motion sensors were sampling at 1/8 Hz and slower than the power consumption sensors. So, we resampled the filtered power consumption sensor data to 1/8 Hz.

### 3.3 Activity probability maps

An activity probability map contains the probability of having a sensor event for each hour of the day. The probability an event of sensor  $i$  occurring in hour  $h$  was computed by  $p(s_{h,i}) = \#s_{h,i} / \#s_i$ , where  $s_i$  was the sensor of interest,  $\#s_{h,i}$  the number of events in hour  $h$  of sensor  $i$ , and  $\#s_i$  the total number of sensor events of sensor  $i$  for the current day. The maps were created for each room of the flat separately and using a subset of all sensors. Figure 2 shows an activity probability map of a kitchen and Fig. 3 of a living room. We established average activity probability maps for each room as baseline for our analysis. The baseline map was the average probability of the maps of 7 days to capture each weekday once. We carefully checked the diaries and none of the participants had any medical condition or health related incidents in the first 7 days.

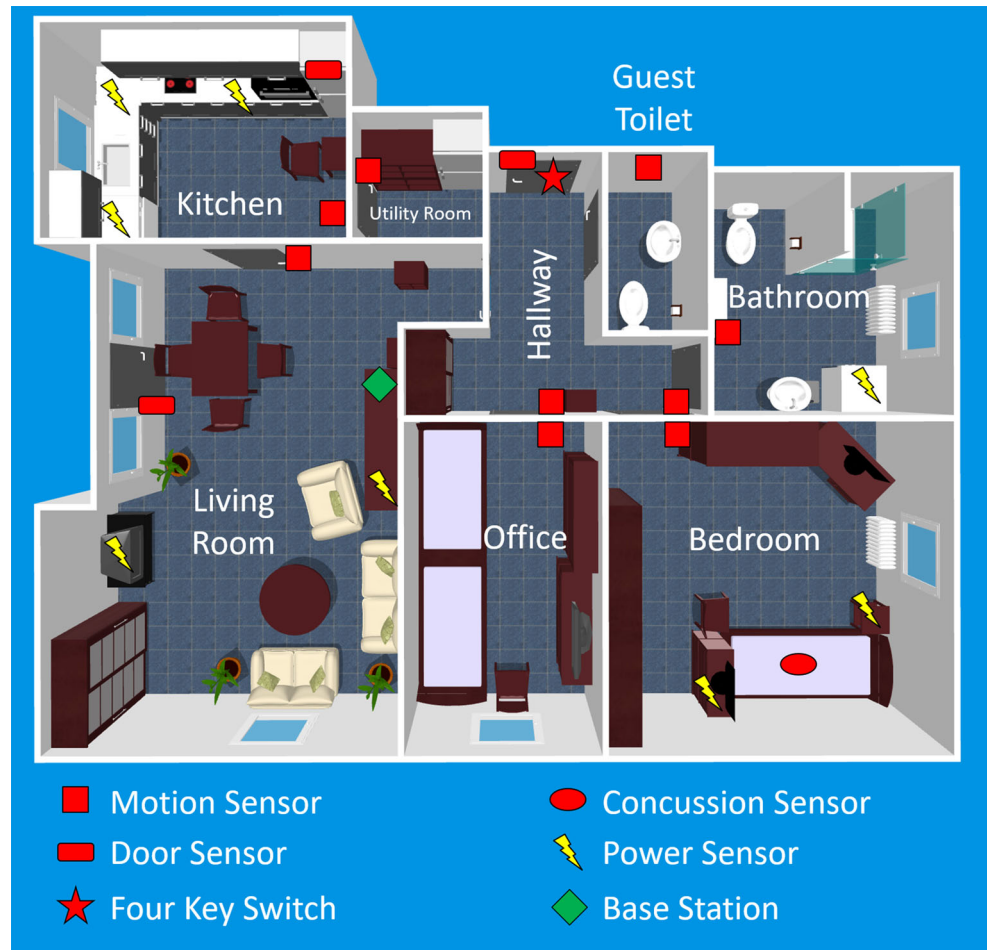
**Table 1** The baseline characteristics of the cohort

	n=20 (m=3,f=17)	Age(y)	Frailty Index(pts.)	SPPB(pts.)	TUG(s)
Mean		84.75	1.90	5.95	17.87
SD ( $\pm$ )		5.19	0.72	2.33	5.33
Range (min-max)		76.00-92.00	1.00-3.00	3.00-11.00	11.16-31.63

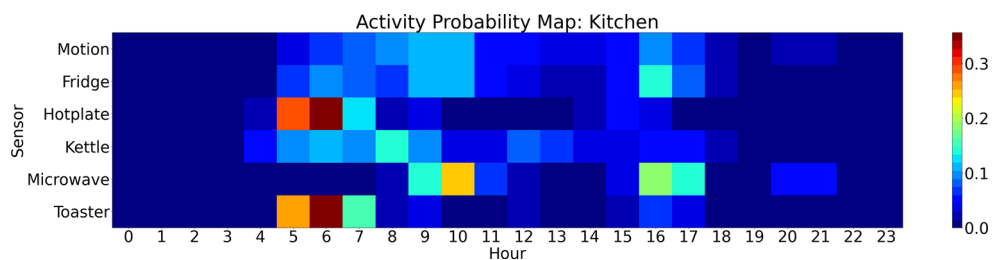
**Table 2** The characteristics of the cohort at the end of the study

n=18 (m=3,f=15)	Age(y)	Frailty Index(pts.)	SPPB(pts.)	TUG(s)
Mean	85.44	2.00	6.61	16.12
SD (±)	4.92	0.97	2.85	5.85
Range (min–max)	77.00–93.00	0.00–4.00	2.00–12.00	8.15–30.06

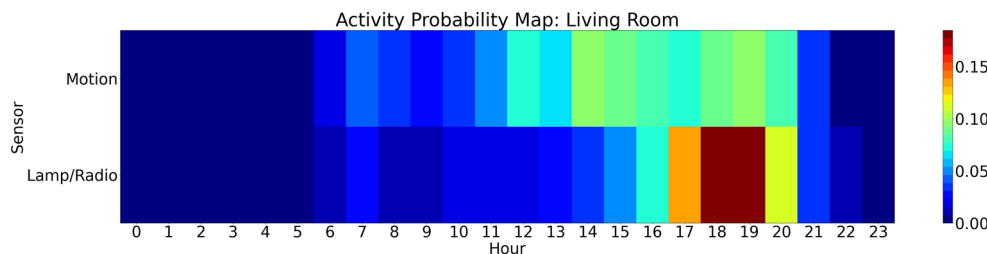
**Fig. 1** A flat of a participant of the OTAGO study



**Fig. 2** One example of an activity probability map of a kitchen. Hours 5 and 6 show high probabilities for using the toaster, the hot plate and the kettle



**Fig. 3** One example of an activity probability map of a living room. The power consumption sensor was attached to a multi socket where a lamp and a radio were connected to



### 3.4 Feature engineering

The feature that was used for training the VAE was the statistical distance measure  $D_{KL}$ . The  $D_{KL}$  is defined as follows

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log \left( \frac{P(x)}{Q(x)} \right) \tag{1}$$

where  $X$  is the probability space, and  $P$  and  $Q$  probability distributions defined over  $X$ . Here  $P$  and  $Q$  were activity probability maps. In case the quotient becomes 0, using

$$\lim_{x \rightarrow 0^+} x \log(x) = 0 \tag{2}$$

results in a contribution of 0 to the sum, if  $P(x) = 0$ . We computed the  $D_{KL}$  for each sensor of the activity probability map and the baseline. The final distance was the average over all  $D_{KL}$  of all rooms. The final feature set was scaled to the interval  $[0, 1]$ . One benefit of the  $D_{KL}$  is its robustness against small probability changes. A small change in the probability map leads to a small change of the  $D_{KL}$ .

### 3.5 Unsupervised statistical concept drift detection

Concept drift detection is a method for detecting drifts and abnormalities in data streams and time series data. The basic idea is to compare new data to old data using a similarity measure and train a classifier. Concept drift detection can be used according to the supervised learning or the unsupervised learning paradigm. In supervised concept drift detection it is assumed that the label is available right after the new data arrived and then based on the label the classifier is updated. However, in case the label is not available, unsupervised learning must be used. In the unsupervised way the decision whether a drift was detected is usually based on a user defined threshold. If the similarity measure exceeds the threshold a drift detection alert is raised. We adapted the general process to our approach and call it *unsupervised statistical concept drift detection*. Figure 4 illustrates the process of concept drift detection.

In the beginning we established a baseline activity probability map as described in Section 3.3 and computed the corresponding features (Section 3.4) to generate the baseline feature set. Then 100,000 samples for training and 50,000 samples for validation are drawn from a normal distribution

$\mathcal{N}(\bar{\mu}, \bar{\sigma}^2)$  parameterised by the sample mean and sample variance. The VAE is trained and validated on the sampled data. The prior, the latent space and posterior were normal distributions, because the CLT establishes that the limiting distribution of a set of random samples is a normal distribution. In the next step, *Concept Drift Detection*, the hypothesis  $H_0$ : *The new sample is not a drift* is tested. The cumulative distribution function of the posterior normal distribution given the new sample  $s$  is evaluated and if  $P(X \leq s) \leq \sigma$  the hypothesis is accepted and rejected otherwise. If  $H_0$  is accepted, the new sample  $s$  is added to the baseline feature set and hence the baseline feature set is updated and the process starts over for the next sample. If the hypothesis is rejected an alarm for a detected drift is raised.

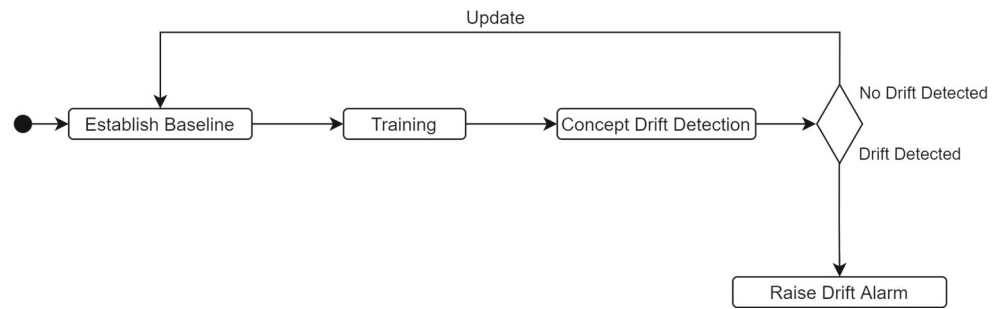
### 3.6 Central limit theorem

The CLT establishes that the limiting random distribution of a sequence of identical independent distributed (i.i.d.) random variables is a standard normal distribution. The samples must be drawn from a distribution with finite mean and finite variance. The mathematical formulation of the Lindeberg–Levy CLT is stated as follows: Let  $\{X_1, \dots, X_n\}$  be a sequence of i.i.d. random variables with  $\mathbb{E}[X_i] = \mu$  and  $\text{Var}[X_i] = \sigma^2 < \infty, i = 1, \dots, n$ . Then  $\lim_{n \rightarrow \infty} \sqrt{n}(\bar{X}_n - \mu) = \mathcal{N}(0, \sigma^2)$  In general,  $n$  must be a sufficient large number, but the CLT itself does not make a statement about the size of  $n$ . Since our sample size  $n$  is limited we used an VAE to approximate the limiting distribution stated by the CLT with a small  $n$ .

### 3.7 Variational autoencoder

An autoencoder is a special type of artificial neural network that learns to efficiently encode data in a latent space. The network is comprised of an encoder and a decoder. The encoder is mapping the input data to a latent space and the decoder is reconstructing the data. The autoencoder is trained in an unsupervised manner by comparing the input data and the corresponding decoded data using a similarity measure. The autoencoder is learning a latent representation of the data. A VAE is a special type of autoencoder and learns the distribution of input data and hence encodes the

**Fig. 4** The illustration of our unsupervised statistical concept drift detection approach



data as distribution over the latent space. For achieving that the latent space is regularised during the training process. The training process has four steps. The first step is to encode the input as a distribution over the latent space. The second step is to draw a sample  $z$  from the latent distribution. In the third step the sample is decoded and the reconstruction error is computed. In the last step, the network is updated by backpropagating the reconstruction error. The latent space must be reparameterised to enable the gradient updating the encoder [36]. Figure 5 illustrates the used VAE architecture.

The encoder learns the latent distribution  $q_\phi(z|x)$  of the posterior distribution  $p_\theta(z|x)$ , where it is assumed that the data is generated by a distribution  $p_\theta(x|z)$ . Here,  $\phi$  and  $\theta$  were the parameters of the encoder and the decoder respectively. In our case the prior distribution was  $\mathcal{N}(\mu_e, \sigma_e)$ , where  $\mu_e$  and  $\sigma_e$  were the sample mean and sample variance of our data. The resultant distributions are

$$q_\phi(z|x) = \mathcal{N}(\mu_{enc}(x), \sigma_{enc}^2(x)) \tag{3}$$

and

$$p_\theta(x|z) = \mathcal{N}(\mu_{dec}(z), \sigma_{dec}^2(z)) \tag{4}$$

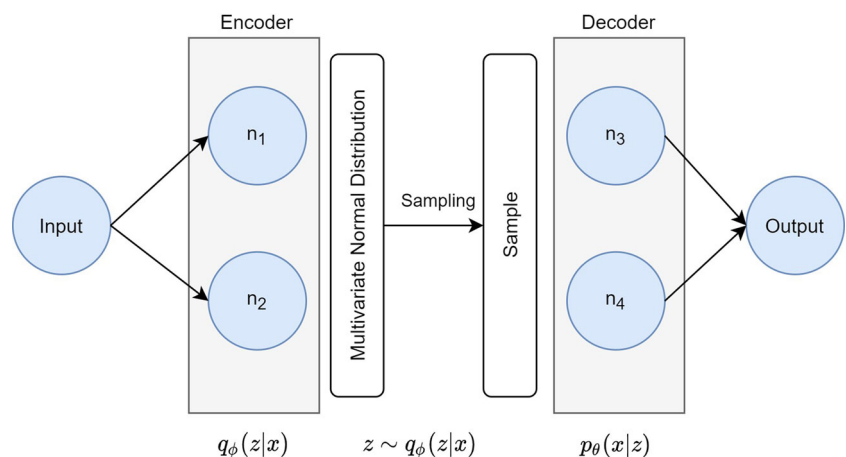
where  $\mu_{enc}(x)$  and  $\sigma_{enc}^2(x)$  are encoder outputs and  $\mu_{dec}(z)$  and  $\sigma_{dec}^2(z)$  the decoder outputs. For evaluating  $H_0$  the cumulative distribution function of  $\mathcal{N}(\mu_{dec}(z), \sigma_{dec}^2(z))$

was evaluated for  $x$ . We trained the VAE using the Evidence Lower Bound Objective (ELBO).

### 3.8 Method validation

We validated our methodology in two steps. The first step was to validate whether the VAE is able to achieve a better approximation of the mean and the variance of the unknown underlying normal distribution than the sample mean and the sample variance. In the second step we validated whether our approach can capture the four common concept drift types. In practical applications the random distribution of the measurements is mostly unknown and must be approximated. The CLT shows that the limiting distribution  $Z$  for random samples  $X_1, \dots, X_n$  with mean  $\mu$  and finite  $\sigma^2$  is the standard normal distribution. The more measurements are available, the better the approximation. We used a standard normal distribution  $\mathcal{N}(0, 1)$  as ground truth and drew two samples. Then we computed the difference between the approximated means and variances and the ground truth 1,000 times. We took the sample mean and the sample variance of all distances to judge the approximation quality of both approaches. To compute the mean and the variance of the VAE approximation we evaluated the VAE at a randomly sampled point from the approximated distribution  $\mathcal{N}(\bar{\mu}, \bar{\sigma}^2)$ . To validate whether our method is able to detect concept drifts we created artificial datasets.

**Fig. 5** The illustration of our VAE. The used activation function was the sigmoid function and the used optimiser was ADAM with an initial learning rate of  $1 \times 10^{-4}$  and an exponential decay after the 5th epoch. The training was automatically stopped, if no improvement of the validation loss larger than 0.001 was found within 5 epochs



We created one dataset for the four typical concept drift types, sudden, gradual, incremental, and recurring described in [37]. Figure 6 illustrates the four different drift types.

Each set contained 300 samples on the domain [0, 1] where 0 denotes no drift and a value larger than 0 a drift. The sudden concept drift data set contained 150 normal samples and 150 drifted samples. The drift occurred in the middle of the series after 150 samples. The gradual drift set contained 165 normal samples and 135 drifted samples. After 10 samples  $n$  drifted samples were inserted;  $n$  counts from 1 to 9. The incremental drift set contained 100 normal and 200 drifted samples. The drift started after sample 100 and was linear incremented until sample 200. The recurring set contained 200 normal samples and 100 drifted samples. The samples between sample 100 and 200 were drifted samples.

## 4 Results

### 4.1 Underlying distribution approximation

Table 3 shows the results of the experiments for validating the VAE approximation of the underlying random distribution.

The average difference of 1,000 runs using two samples was 0.91 and the variance was 0.41, and the average difference using the VAE approximation was 0.73 and the variance 0.33. The results show that the VAE approximates the unknown underlying distribution more accurate than the sample mean and the sample variance.

**Table 3** The results of the distribution approximation. The mean and variance rows show the mean and variance of the difference between the baseline {0, 1} and the approximations

	Sampled	Approximation
Mean	0.91	0.73
Variance	0.41	0.33

### 4.2 Artificial dataset

The results of our second validation step are shown in Table 4.

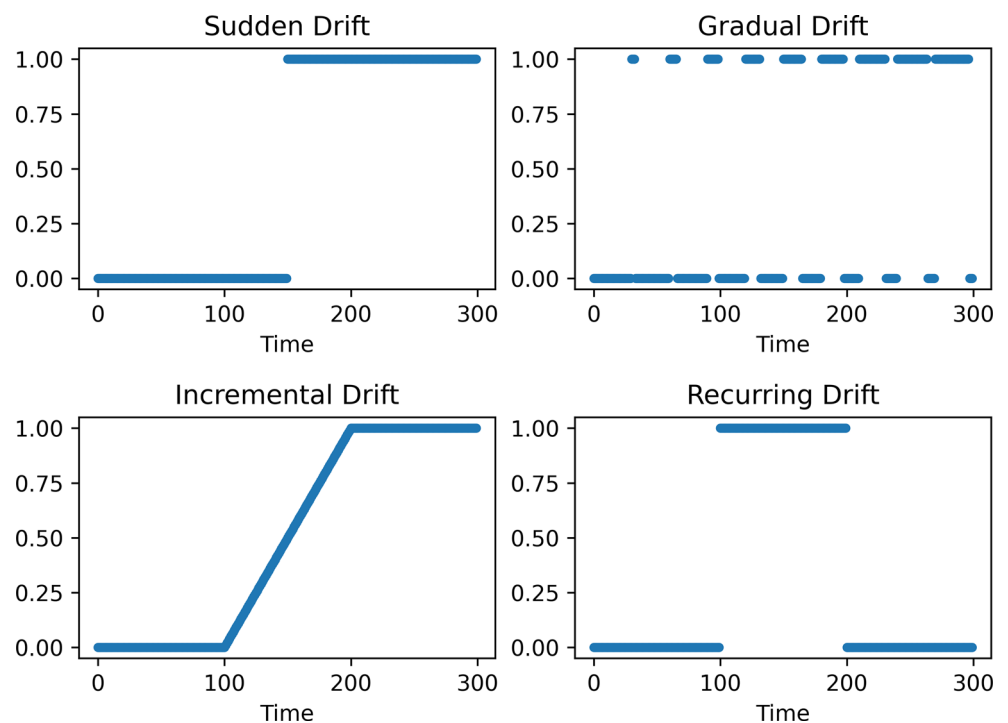
The best scores were 1.0 on the sudden drift set and the recurring drift set. The worst scores of 0.91, 0.89, and 0.83 of the F<sub>1</sub>-Score, accuracy, and recall were on the incremental drift set. The results on the gradual drift set were a F<sub>1</sub>-Score of 0.98, an accuracy of 0.99, and a recall of 0.97. All the scores are referring to the artificial dataset. The lack of labels prevented calculating scores on our real-world dataset.

### 4.3 Real-world dataset

Figures 7 and 8 show the distributions parameterised by the sample mean and sample variance and parameterised by the parameters estimated by the VAE. The red histogram is the baseline distribution of 7d and the blue histogram is the final distribution after processing all data.

The sample mean and the sample variance at baseline for participant 1 were 3.56 and 3.59. The mean and variance

**Fig. 6** The four typical drift types we used for validating our method



**Table 4** The results of the concept drift validation

	Sudden	Gradual	Incremental	Recurring
F <sub>1</sub> -Score	1.0	0.98	0.91	1.0
Accuracy	1.0	0.99	0.89	1.0
Precision	1.0	1.0	1.0	1.0
Recall	1.0	0.97	0.83	1.0

estimated by the VAE were 3.52 and 1.04. At the end the sample mean and the sample variance were 2.60 and 0.73. The final estimated mean and variance were 1.85 and 0.40.

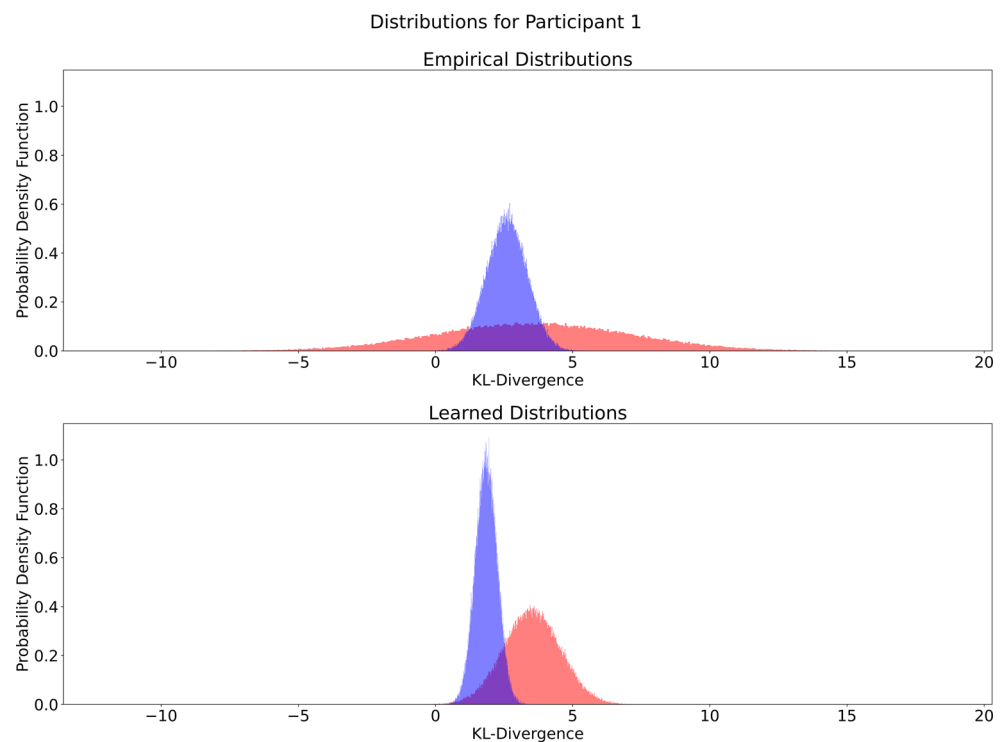
After the first 66 days of the study participant 16 started a chemotherapy and afterwards no normal days were found anymore. The sample mean and variance at the beginning of the study were 2.06 and 0.47. The mean and variance estimated by the VAE were 2.86 and 0.31. The sample mean and variance at the end of the study were 1.81 and 0.46. The VAE estimation of the mean and variance at the end were 1.51 and 0.21. Table 5 shows the number of normal and abnormal days for different  $\sigma$  for each participant.

## 5 Discussion

The results on the artificial concept drift datasets and the VAE approximation were easy to judge, because ground truth values were available. Our real-world dataset did not have reliable ground truth values. First of all, we could see that the algorithm behaves as expected, when adjusting the

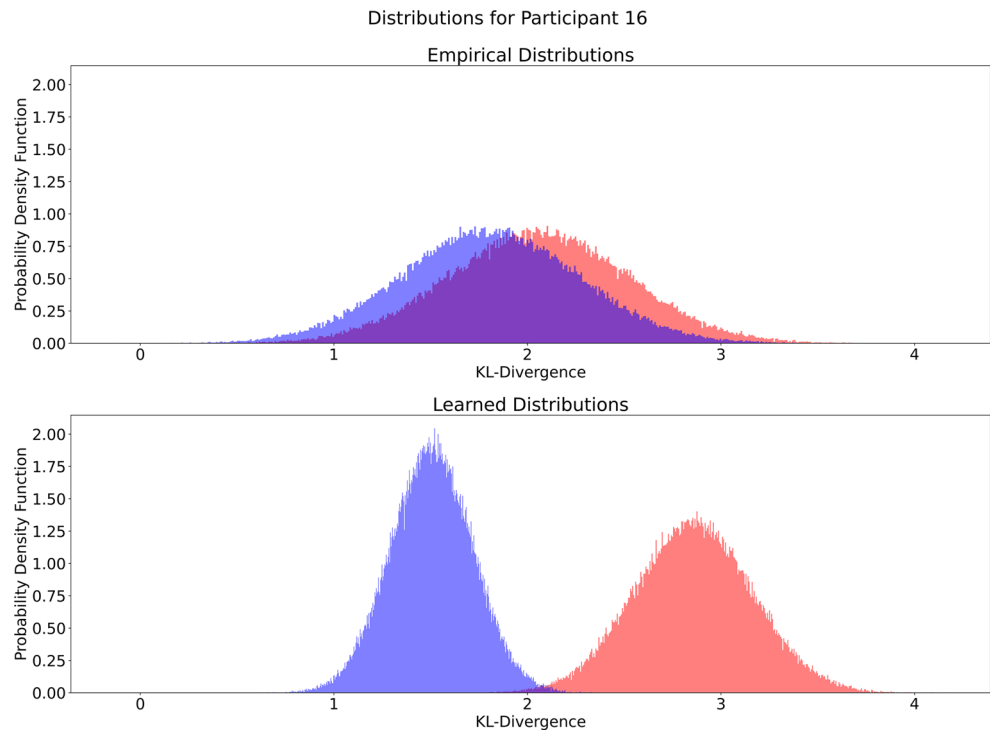
factor of  $\sigma$ . The larger the factor of  $\sigma$ , the more a value can deviate from the mean without being considered as abnormal. The results showed as well that our method is able to adapt to the volatile behaviour of humans. A good indicator for the suitability of our method was participant 16. The participant got cancer during the study and had to undergo chemotherapy. The participant deceased before the study finished. Chemotherapy has severe side effects and is stressful for body and mind. Choosing  $1\sigma$  our method did not detect any normal day after the first chemotherapy treatment. Our approach was able to capture the change much better than the common statistical approach, because the shift of the mean was much larger. Another interesting case was participant 11. After a fall incident the participant had an injured ankle and wore an orthosis. The first 11 days were abnormal according to our algorithm, but then after the 11th day normal days were found again. Perhaps the participant adjusted to wearing the orthosis and the behaviour normalised again. Other participants were hospitalised during the study and those days were abnormal as well. Participant 3 lend the flat to relatives while on vacation, so all 5 days were recognised as abnormal. With  $1\sigma$  our method considers all days where the participants had a medical condition as abnormal, but days where visitors came or the participants went on trips were considered as abnormal as well. One participant reported that a motion sensor fell off and all days until it was fixed by staff of the university were considered as abnormal. We found that all days participant 15 had an inflammation of the bladder and participant 16

**Fig. 7** The drift of the distributions of participant 1. All severe incidents are captured using  $1\sigma$  for testing the  $H_0$ . The blue histograms show the distributions at the beginning and the red at the end of the study





**Fig. 8** The drift of the distributions of participant 16. After beginning the chemotherapy no normal day is found using  $1\sigma$  for testing the  $H_0$ . The blue histograms show the distributions at the beginning and the red at the end of the study



had gastroenteritis were considered as abnormal. Even though the diaries were an unreliable ground truth source, we considered severe medical conditions like falls, bone fractures,

**Table 5** The results of our statistical concept drift detection algorithm on the real-world dataset for different  $\sigma$ . normal days / abnormal days

ID	$1\sigma$	$1.5\sigma$	$2\sigma$
1	274/32	299/7	304/2
2	171/130	272/29	294/7
3	241/94	316/19	327/8
4	265/39	292/12	300/4
5	189/141	311/19	327/3
6	212/90	271/31	297/5
7	248/79	318/9	323/4
8	95/56	139/12	148/3
9	233/30	257/6	260/3
10	293/15	304/4	307/1
11	214/101	282/33	309/6
12	275/81	346/10	356/0
13	256/54	293/17	302/8
14	177/31	197/11	205/3
15	187/107	279/15	293/1
16	110/42	145/7	149/3
17	284/15	295/4	296/3
18	272/61	319/14	330/3
19	294/43	300/37	332/5
20	221/70	278/13	286/5

hospitalisation, gastroenteritis, inflammation of the bladder, and chemotherapy as reliable. Other information like going out for a walk, having visitors, and playing games were not considered as reliable information. The algorithm considered days as abnormal where no reliable information is available. That is an indicator that our method has false positives. However, in the medical context a false positive is preferable over a false negative. Our system keeps a human in the loop and in case of an alarm the reason can be found. The combination our unsupervised concept drift detection method with a high precision and an medical expert with a high recall may lead to a high performance. There were the factor of  $\sigma$  and the days of the baseline to choose. The baseline of 7 days were chosen based on the assumption that routines on day scale are dependent on the day of the week. That may not be necessarily true. In clinical application the initial baseline would be 2 days because at least 2 days are needed to calculate the variance. Since the label would be available only normal days can be considered for the baseline. The factor of  $\sigma$  would be adapted in online learning fashion. The parameter controls the sensitivity of the model. In case of many false alarms the factor would be increased.

## 6 Conclusions

We introduced our novel unsupervised statistical concept drift detection method, and the results on the artificial dataset showed that our method was able to detect concept

drifts using activity probability maps derived from home automation data and the  $D_{KL}$ . Moreover, we showed that our new approach using VAE for estimating the probability distribution is more accurate than the sample mean and sample variance. We evaluated our method on a real-world dataset of 20 (pre-)frail older adults and showed that our method was successful in adapting to a change in behaviour. Moreover, we showed that our method also detects abnormal behaviour in case of medical condition. Except for the medical conditions no reliable ground truth labels were available and so there may be false positives. Our approach was designed with humanised and personalised care in mind and keeps a human in the loop. In this context our approach is supposed to have a high precision and a high recall. The next is to test our approach on a similar dataset with reliable ground truth values. Considering the medical perspective a clinical study designed to evaluate our approach would be the most sophisticated way of testing our method.

**Acknowledgements** We acknowledge Enno-Edzard Steen (University of Oldenburg) for setting up and maintaining the sensor systems in the flats, Bianca Sahlmann (University of Oldenburg), and Lena Elgert (Peter L. Reichertz Institut) for performing the assessments. The experiments were performed at the HPC Cluster CARL, located at the University of Oldenburg (Germany) and funded by the DFG through its Major Research Instrumentation Programme (INST 184/157-1 FUGG) and the Ministry of Science and Culture (MWK) of the Lower Saxony State. Moreover, we gratefully acknowledge the support of the NVIDIA Corporation with the donation of the TITAN V GPU used for this research.

**Author Contributions** Conceptualization, B.F., T.S. and A.H.; methodology, B.F.; software, B.F.; validation, B.F. and T.S.; formal analysis, B.F.; investigation, B.F. and T.S.; resources, A.H.; data curation, B.F.; writing—original draft preparation, B.F. and T.S.; writing—review and editing, B.F., T.S. and A.H.; visualization, B.F.; supervision, A.H.; project administration, A.H.; funding acquisition, A.H. All authors have read and agreed to the published version of the manuscript.

**Funding** Open Access funding enabled and organized by Projekt DEAL. This research received no external funding.

**Availability of data and materials** The data is not publicly available due to privacy concerns.

## Declarations

**Conflict of Interests** The authors declare no conflict of interest.

**Informed Consent** Informed consent was obtained from all subjects involved in the study.

**Institutional Review Board** The OTAGO study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of Carl von Ossietzky University (protocol code: Drs.27/2014, date of approval: 30.04.2014).

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Galvin JE, Sadowsky CH (2012) Practical guidelines for the recognition and diagnosis of dementia. *J Am Board Fam Med* 25(3):367–382. <https://doi.org/10.3122/jabfm.2012.03.100181>
- Gerlach LB, Kales HC (2018) Managing behavioral and psychological symptoms of dementia. *Psychiatr Clin N Am* 41(1):127–139. <https://doi.org/10.1016/j.psc.2017.10.010>, *Geriatric Psychiatry*
- Gerlach LB, Kales HC (2018) Behavioral problems and dementia. *Clin Geriatr Med* 34(4):637–651. <https://doi.org/10.1016/j.cger.2018.06.009>
- Feast A, Orrell M, Charlesworth G, Melunsky N, Poland F, Moniz-Cook E (2016) Behavioural and psychological symptoms in dementia and the challenges for family carers: Systematic review. *Br J Psychiatr* 208(5):429–434. <https://doi.org/10.1192/bjp.bp.114.153684>
- Chen H, Jiang M, Liu Y, He J, Li H (2020) Review on machine learning and its application in atmospheric science and human behavior recognition. In: *Proceedings of the 2020 3rd International Conference on Signal Processing and Machine Learning*. SPML 2020. Association for Computing Machinery, New York, NY, USA, pp 98–104. <https://doi.org/10.1145/3432291.3432311>
- Kullback S (1959) *Information theory and statistics*. John Wiley & Sons
- Singh D, Merdivan E, Psychoula I, Kropf J, Hanke S, Geist M, Holzinger A (2017) Human activity recognition using recurrent neural networks. In: Holzinger A, Kieseberg P, Tjoa AM, Weippl E (eds) *Machine Learning and Knowledge Extraction*. Springer International Publishing, Cham, pp 267–274
- Liciotti D, Bernardini M, Romeo L, Frontoni E (2019) A sequential deep learning application for recognising human activities in smart homes, vol 396. <https://doi.org/10.1016/j.neucom.2018.10.104>
- Kolkar R, Geetha V (2021) Human activity recognition in smart home using deep learning techniques. In: *2021 13th International conference on information communication technology and system (ICTS)*, pp 230–234
- Xuan J, Lu J, Zhang G (November 2020) Bayesian nonparametric unsupervised concept drift detection for data stream mining. *ACM Trans. Intell. Syst. Technol.* 12:1. <https://doi.org/10.1145/3420034>
- dos Reis DM, Flach P, Matwin S, Batista G (2016) Fast unsupervised online drift detection using incremental kolmogorov-smirnov test *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. Association for Computing Machinery, New York, NY, USA, pp 1545–1554. <https://doi.org/10.1145/2939672.2939836>
- Dahmen J, Cook DJ (2021) Indirectly supervised anomaly detection of clinically meaningful health events from smart home

- data. *ACM Trans Intell Syst Technol* 12:2. <https://doi.org/10.1145/3439870>
13. Liu A, Lu J, Liu F, Zhang G (2018) Accumulating regional density dissimilarity for concept drift detection in data streams. *Pattern Recogn* 76:256–272. <https://doi.org/10.1016/j.patcog.2017.11.009>, <https://www.sciencedirect.com/science/article/pii/S0031320317304636>
  14. Cerqueira V, Gomes HM, Bifet A (2020) Unsupervised concept drift detection using a student–teacher approach. In: Appice A, Tsoumakas G, Manolopoulos Y, Matwin S (eds) *Discovery Science*. Springer International Publishing, Cham, pp 190–204
  15. Li B, Wang Y, Yang D, Li Y, Ma X (2019) Faad: an unsupervised fast and accurate anomaly detection method for a multi-dimensional sequence over data stream. *Front Inf Technol Electron Eng* 20:388–404. <https://doi.org/10.1631/FITEE.1800038>
  16. Eisa S, Moreira A (2017) A behaviour monitoring system (bms) for ambient assisted living. *Sensors* 17:9. <https://doi.org/10.3390/s17091946>, <https://www.mdpi.com/1424-8220/17/9/1946>
  17. Azefack C, Phan R, Augusto V, Gardin G, Coquard C, Bouvier R, Xie X (2019) An approach for behavioral drift detection in a smart home
  18. Sfar H, Bouzeghoub A, Raddaoui B (2018) Early anomaly detection in smart home: A causal association rule-based approach. *Artif Intell Med* 91:57–71. <https://doi.org/10.1016/j.artmed.2018.06.001>, <https://www.sciencedirect.com/science/article/pii/S0933365717305985>
  19. Jakkula V, Cook DJ (2008) Anomaly detection using temporal data mining in a smart home environment. *Methods Inf Med* 47(1):70–75. <https://doi.org/10.3414/me9103>
  20. Jalali N, Sahu KS, Oetomo A, Morita PP (2020) Understanding user behavior through the use of unsupervised anomaly detection: Proof of concept using internet of things smart home thermostat data for improving public health surveillance. *JMIR Mhealth Uhealth* 8(11):e21209. <https://doi.org/10.2196/21209>, <http://mhealth.jmir.org/2020/11/e21209/>
  21. Shang C, Chang C-Y, Chen G, Zhao S, Lin J (2020) Implicit irregularity detection using unsupervised learning on daily behaviors. *IEEE J Biomed Health Inform* 24(1):131–143. <https://doi.org/10.1109/JBHI.2019.2896976>
  22. Diraco G, Leone A, Siciliano P (2019) Ai-based early change detection in smart living environments. *Sensors* 19:16. <https://doi.org/10.3390/s19163549>, <https://www.mdpi.com/1424-8220/19/16/3549>
  23. von Ossietzky Universität Oldenburg C (2020) Otago. <https://uol.de/en/amt/research/projects/otago>, Accessed: 2020-12-20
  24. Searle SD, Mitnitski A, Gahbauer EA, Gill TM, Rockwood K (2008) A standard procedure for creating a frailty index, vol 8. <https://doi.org/10.1186/1471-2318-8-24>
  25. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, Seeman T, Tracy R, Kop WJ, Burke G, McBurnie MA (2001) Frailty in older adults: Evidence for a phenotype. *The J Gerontol: Series A* 56:M146–M157. <https://doi.org/10.1093/gerona/56.3.M146>
  26. Tinetti ME (1986) Performance-oriented assessment of mobility problems in elderly patients. *J Am Geriatr Soc* 34(2):119–126. <https://doi.org/10.1111/j.1532-5415.1986.tb05480.x>
  27. Guralnik JM, Simonsick EM, Ferrucci L, Glynn RJ, Berkman LF, Blazer DG, Scherr PA, Wallace RB (1994) A short physical performance battery assessing lower extremity function: Association with self-reported disability and prediction of mortality and nursing home admission. *J Gerontol* 49:M85–M94. <https://doi.org/10.1093/geronj/49.2.M85>
  28. Podsiadlo D, Richardson S (1991) The timed “up & go”: A test of basic functional mobility for frail elderly persons. *J Am Geriatr Soc* 32:142–148. <https://doi.org/10.1111/j.1532-5415.1991.tb01616.x>
  29. Sayer AA, Kirkwood TBL (2015) Grip strength and mortality: a biomarker of ageing? *Lancet* 386(9990):226–227. [https://doi.org/10.1016/S0140-6736\(14\)62349-7](https://doi.org/10.1016/S0140-6736(14)62349-7)
  30. Lawton MP, Brody EM (1969) Assessment of older people: Self-maintaining and instrumental activities of daily living. *Gerontologist* 9:179–186. [https://doi.org/10.1093/geront/9.3.Part\\_1.179](https://doi.org/10.1093/geront/9.3.Part_1.179)
  31. Herdman M, Gudex C, Lloyd A, Janssen B, Kind P, Parkin D, Bonsel G, Badia X (2011) Development and preliminary testing of the new five-level version of eq-5d (eq-5d-5l). *Qual Life Res* 20:1727–1736. <https://doi.org/10.1007/s11136-011-9903-x>
  32. Vellas B, Guigoz Y, Garry PJ, Nourhashemi F, Bennahum D, Lauque S, Albaredo J-L (1999) The mini nutritional assessment (mna) and its use in grading the nutritional state of elderly patients. *Nutrition* 15:116–122. [https://doi.org/10.1016/S0899-9007\(98\)00171-3](https://doi.org/10.1016/S0899-9007(98)00171-3)
  33. In Motion SD (2019) Shimmer3 wireless sensor platform. [http://www.shimmersensing.com/images/uploads/docs/Shimmer3\\_Spec\\_Sheet\\_V1.8.pdf](http://www.shimmersensing.com/images/uploads/docs/Shimmer3_Spec_Sheet_V1.8.pdf)
  34. Columbus (2011) V-990 multifunction gps data logger user manual. <https://cbgps.com/download/Columbus.V-990.User.Manual.V1.0.ENG.pdf%>
  35. Hassler AP, Menasalvas E, García-García FJ, Rodríguez-Mañas L, Holzinger A (2019) Importance of medical data pre-processing in predictive modeling and risk factor discovery for the frailty syndrome. *BMC Medical Inform Decis Mak* 19(1):33. <https://doi.org/10.1186/s12911-019-0747-6>, <https://doi.org/10.1186/s12911-019-0747-6>
  36. Kingma DP, Welling M (2019) An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning* 12:307–392. <https://doi.org/10.1561/22000000056>
  37. Lemaire V, Salperwyck C, Bondu A (2015) A survey on supervised classification on data streams. [https://doi.org/10.1007/978-3-319-17551-5\\_4](https://doi.org/10.1007/978-3-319-17551-5_4)

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.