



OPEN Tackling inter-subject variability in smartwatch data using factorization models

Arman Naseri^{1,2✉}, David M. J. Tax¹, Ivo van der Bilt² & Marcel Reinders¹

Smartwatches enable longitudinal and continuous data acquisition. This has the potential to remotely monitor (changes) of the health of users. However, differences among subjects (inter-subject variability) limit a model to generalize to unseen subjects. This study focused on binary classification tasks using heart rate and step counter from smartwatches, including night/day and inactive/active classification, as well as sleep and SpO₂-related (oxygen saturation) tasks. To address inter-subject variability, we explored different transforming and normalization regimes for time series including per-subject and population-based strategies. We propose a modified factorized autoencoder, which separates the data into two latent spaces capturing class-specific and subject-specific information. Our proposed generalized factorized autoencoder and triplet factorized autoencoder improved classification accuracy over the baseline from 74.8 (\pm 10.5) to 83.1 (\pm 5.1) and 83.4 (\pm 5.3), respectively, for night/day classification, gains for inactive/active classification were modest, improving from 84.3 (\pm 9.4) to 86.9 (\pm 4.4) and 86.6 (\pm 4.3), respectively. Our study highlights challenges of handling inter-subject variability in smartwatch data and how factorization models can be used to enable more robust and personalized health monitoring solutions for diverse populations.

Keywords Inter-subject variability, Machine learning, Smartwatch, Neural networks

Consumer-grade smartwatches have transformed the landscape of personal health monitoring by enabling the continuous recording of vital signals, such as heart rate, physical activity, and sleep patterns. Originally developed for recreational use and fitness tracking, these devices have rapidly gained attention in the medical domain due to their widespread availability, non-invasiveness, and relatively low cost. When combined with machine learning (ML) techniques, smartwatch data has demonstrated significant potential for early detection and monitoring of various health conditions, including cardiovascular diseases, diabetes, and respiratory disorders^{1,2}. This shift marks an important step toward integrating wearable technologies into predictive and preventive healthcare.

In the context of cardiovascular disease, early detection and monitoring are particularly challenging due to the subtle, gradual changes in cardiovascular signals that precede acute events³. Cardiovascular disease affects millions worldwide, with significant mortality and morbidity rates, and is a leading cause of death^{4,5}. Traditional diagnostic methods rely on echocardiography and clinical biomarkers^{6,7}, which are often inaccessible for continuous monitoring. Wearable devices provide an opportunity to bridge this gap by enabling remote, real-time monitoring of early warning signs such as changes in heart rate dynamics, physical activity, and sleep patterns. However, the successful implementation of such systems requires overcoming challenges related to variability and data quality.

Cardiovascular properties, including heart rate dynamics, are highly individualized and influenced by various factors. Studies have shown that body weight⁸, sex⁹, age, and height^{10,11} play significant roles. Correspondingly, heart size and function scale with body mass index and exhibit sex-specific differences in cardiac output and structure. Furthermore, longitudinal studies using consumer-grade wearables like Fitbit have shown that resting heart rate and heart rate variability exhibit substantial inter-subject (between individuals) and intra-subject (within individuals over time) variability¹². Such variability is further influenced by lifestyle factors, environmental conditions, and emotional states, complicating the interpretation of physiological signals for population-wide predictions.

Despite limitations, such as low sampling rates and reliance on derived metrics (e.g., estimated heart rate variability), smartwatches provide surprisingly detailed insights into physiological states. For example, studies have demonstrated their ability to capture heart rate patterns that correlate with respiratory infections¹³ or cardiac conditions like atrial fibrillation and heart failure^{14–16}. However, this level of detail also introduces challenges

¹Delft University of Technology, Delft, The Netherlands. ²Haga Teaching Hospital, The Hague, The Netherlands. ✉email: a.naserijahfari@hagaziekenhuis.nl

in ML-based analyses. Measurements of heart rate often encode individual-specific characteristics, which may dominate the patterns learned by ML models, leading to poor generalization across unseen subjects. This issue arises when subject-specific variability (e.g., baseline heart rate differences) confounds the relationship between the physiological signal and the predictive task. Such confounding can degrade the performance of ML models, as they may focus on irrelevant patterns rather than generalizable features predictive of health outcomes.

Addressing inter- and intra-subject variability in smartwatch data has been a major focus in recent literature. Techniques such as domain adaptation¹⁷, domain generalization¹⁸, batch effect correction¹⁹, and feature normalization²⁰ have been explored to improve generalization. Domain invariance techniques, including adversarial training and representation learning, aim to disentangle subject-specific factors from class-relevant signals^{21–23}. These strategies are essential for creating robust predictive models that can generalize across diverse populations, a critical requirement for clinical applications of wearable data.

In this study, we propose a novel machine learning framework for predicting several tasks such as night/day and active/inactive binary classification, based on heart rate and step counter data obtained from smartwatches. Such tasks can for example help in assessing circadian rhythm disruptions or monitoring recovery post-surgery, which are important indicators of overall health. Our approach incorporates methods to address inter- and intra-subject variability, including normalization techniques, domain-invariant feature extraction and an extended loss function, to enhance the model's generalizability. By leveraging the continuous and rich data captured by consumer-grade wearables, we aim to provide a scalable solution for smartwatch monitoring, advancing the integration of wearable technology into cardiovascular healthcare.

Methods

A general overview of our machine learning approach is given in Fig. 1. Roughly, it consists of three steps. First, the smartwatch time series data is transformed to capture the dynamics and segmented into windows of several minutes. Then the data is normalized for which we propose and analyze several approaches (either population-based or subject-based). Lastly, the windows are fed into a factorized autoencoder that is specifically designed to disentangle variations between subjects and variations between classes by mapping the samples simultaneously in a so-called domain-space (representing differences between subjects) as well as a class-space (representing differences between the classes). For that, we do propose three different variants based on a contrastive technique, two of which are use our generalized factorized loss and one is based on our proposed triplet factorized loss. The following gives more details on each of the steps.

Transformation

We apply two transformations to the heart rate time series to capture dynamic changes while minimizing the influence of individual baseline variability. Heart acceleration ($acc(t)$) is calculated as the first derivative of the heart rate hr , defined as the difference between the heart rate at time t and $t - 1$ (Eq. 1). This metric isolates the temporal rate of change in heart rate, making it insensitive to baseline heart rate differences, which vary significantly across subjects and may confound comparisons. Additionally, we compute the normalized (relative) heart acceleration ($acc_{rel}(t)$) as the ratio of heart acceleration to the heart rate at time t (Eq. 2). This normalization accounts for variability in heart rate magnitude, enabling more consistent interpretation of changes across different heart rate levels:

$$acc(t) = hr(t) - hr(t - 1) \quad (1)$$

$$acc_{rel}(t) = \frac{hr(t) - hr(t - 1)}{hr(t)} \quad (2)$$

Segmentation

The (transformed) heart rate timeseries is segmented using a sliding window of 240 samples (20 min) with a stride of 36 samples (3 min), at a sampling rate of once per 5 s. This window length was selected after hyperparameter tuning which included windows of length 120 (10 min), 360 (30 min) and 720 (60 min). To ensure data quality, windows containing timegaps of larger than 5 s, indicative of missing values are excluded. Each window is then assigned a label according to the criteria in Table 1. An SpO2 lower than 95% is generally undesirable²⁴, but as a margin to distinguish abnormal events from normal ones, we label SpO2 levels lower than 90% as abnormal.

Normalization

We aim to further reduce inter-subject variability in heart rate data through the application of multiple techniques based on Z-normalization. Our approach leverages multiple variations of Z-score normalization tailored to different subsets of the data to enhance the consistency and comparability of heart rate measurements across individuals while accounting for subject-specific characteristics.

For population-based normalization, the heart rate time series of all training samples across all training subjects are aggregated to compute the population mean and standard deviation. The heart rate values are then standardized by subtracting the population mean and scaling them to unit variance. The calculated population mean and standard deviation are subsequently applied to normalize the heart rate time series of test subjects.

For per-subject normalization, Z-score normalization is applied individually to each subject. Two approaches are considered (and illustrated in Fig. 1B): In the *per-test subject normalization*, after training with the population-based normalization, an initial portion of each test subject's data—up to the first 60% of their time series—is used to calculate a subject-specific mean and variance. The remaining test data is then normalized using these parameters. This approach ensures that each subject's data is standardized based on their unique heart rate characteristics. However, it requires a “burn-in” period to gather sufficient data for effective calibration.

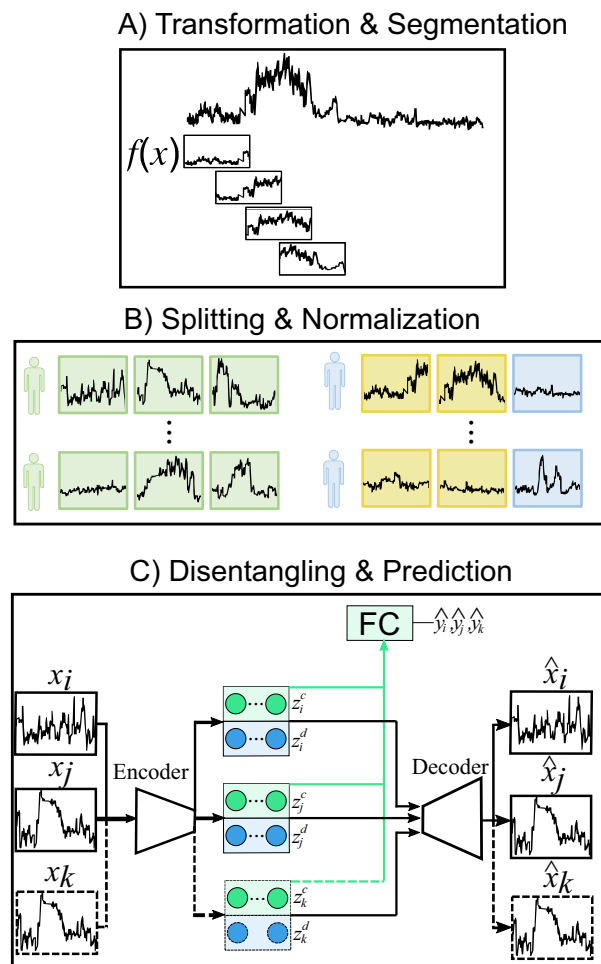


Fig. 1. Overview of the steps of the proposed machine learning approach. **(A)** The timeseries signal is transformed and segmented using a sliding window. **(B)** Subjects are split as train (green, left) or test (blue, right) subjects. Windows of test subjects are further split in calibration (yellow) and test (blue) windows. **(C)** Windows are fed in factorized autoencoder models (three at a time for the triplet factorized autoencoder), where x_i, x_j refer to windows from the same subject with the same class label, and x_k refers to a window of a different subject and class label. The corresponding loss function consists of three main components (Eq. 3). A fully-connected (FC) layer uses only the class latent space z^c to predict the class label to determine the cross-entropy loss. Both the domain latent space z^d and class latent space z^c are fed into the decoder to reconstruct the original input windows to determine the reconstruction loss. Finally, the class latent space and domain latent space are optimized using either our generalized factorized loss or triplet factorized loss, as described in “Factorized autoencoders” section.

Task	Class 1	Class 2
Night/day	Timestamp between 00:00 and 05:00	timestamp between 11:00 and 18:00
Inactive/active	Average steps < 5 per minute	Average steps > 50 per minute
SpO2 (ab)normal	Average SpO2 90	Average SpO2 > 95
Light/deep sleep	Light sleep	Deep sleep
Light/restless sleep	Light sleep	Restless sleep
Light/REM sleep	Light sleep	REM sleep

Table 1. Tasks (rows) and corresponding derived class labels (columns).

Note that this only affects the model during the test phase. For the *per-train subject normalization* the Z-score normalization is performed separately for each training subject, using their individual mean and variance, computed on the entire training data of that subject. Note that this only affects the model during the training phase. The test subjects are processed similarly to the per-test subject scenario.

In both the population-based and per-subject normalization approaches, we explore two variations: normalizing based on all heart rate samples or exclusively on heart rate samples corresponding to zero steps (inactivity normalization). The motivation for the latter is to better estimate resting heart rate, as it focuses on periods of inactivity, which are less influenced by diverse and subject-specific physical activities. Normalization during inactive moments is hypothesized to provide a more consistent baseline across subjects, thereby enhancing the effectiveness of the normalization process.

Factorized autoencoders

We investigated methods to mitigate inter-subject variability using contrastive²⁵ and triplet²⁶ similarity learning. Specifically, we adopted the sequential factorized autoencoder (FAE) framework used by Gyawali et al.²¹. This approach addressed inter-subject variability by employing two key components. *Factorized latent representations*: The bottleneck layer of an autoencoder is partitioned into two distinct latent spaces: a class latent space (z^c) and a domain latent space (z^d). *Contrastive loss function*: A contrastive loss is applied to encourage the network to optimize class separability within the class latent space while making it invariant to subject-specific variability. Simultaneously, the domain latent space is trained to capture subject-specific characteristics while remaining independent of class labels. This separation allows the class latent space representation to be effectively used for downstream tasks.

Building on this framework, we generalized the original loss function and propose the Generalized Factorized Autoencoder (GFAE) defined as:

$$L = L^{ce} + \alpha \left(\beta L^c(z_i^c, z_j^c, y_{ij}^c) + (1 - \beta) L^d(z_i^d, z_j^d, y_{ij}^d) \right) + \gamma L^{rec} \quad (3)$$

where L^c is the class loss defined by the similarity between a pair of class latent samples (z_i^c and z_j^c) and the pair's corresponding class label y_{ij}^c . Similarly, L^d is the domain loss between a pair of domain latent samples (z_i^d and z_j^d) and the pair's corresponding domain label y_{ij}^d , where β trades off the class loss and domain loss. L^{ce} and L^{rec} represent the cross-entropy loss and reconstruction loss and are defined as:

$$L^{ce} = -\frac{1}{2} \sum_{l \in \{i,j\}} (y_l \log(\hat{y}_l^c) + (1 - y_l) \log(1 - \hat{y}_l^c)) \quad (4)$$

$$L^{rec} = \frac{1}{2} \sum_{l \in \{i,j\}} \|x_l - \hat{x}_l\|^2 \quad (5)$$

α and γ are the regularization coefficients for the contrastive loss and reconstruction loss, respectively.

The class loss is defined as:

$$L^c(z_i^c, z_j^c, y_{ij}^c) = y_{ij}^c \|z_i^c - z_j^c\|_2^2 + (1 - y_{ij}^c) \max(0, m^c - \|z_i^c - z_j^c\|_2^2) \quad (6)$$

where y_{ij}^c equals one if the samples that form the pair have the same class labels, and zero otherwise. Regardless of which subjects the pair originates from, the first term aims to project samples of the same class close to each other in the class latent space. When the samples do not have the same class labels, they are kept distant of each other up to a threshold defined by the class margin, m^c .

The domain loss is similarly defined as:

$$L^d(z_i^d, z_j^d, y_{ij}^d) = y_{ij}^d \|z_i^d - z_j^d\|_2^2 + (1 - y_{ij}^d) \max(0, m^d - \|z_i^d - z_j^d\|_2^2) \quad (7)$$

where y_{ij}^d equals one if the pair have the same domain labels and zero otherwise. Regardless of which classes the pair originates from, the first term aims to project samples of the same subject close to each other in the domain latent space. When the samples do not have the same domain labels, they are kept distant of each other up to a threshold defined by the domain margin, m^d .

In the original FAE, the parameter β was set to 0.5, equally balancing the class and domain losses. However, depending on the nature of the data, the model may benefit from adjusting this weighting to account for varying levels of inter-subject variability. In particular, the severity of inter-subject differences can influence how much emphasis should be placed on minimizing class versus domain loss.

Additionally, in the original FAE framework, the domain loss did not account for situations where pairwise latent representations originated from different subjects. Specifically, in Eq. (7), y_{ij}^d is always set to 1, effectively deactivating the second term. Explicitly modeling this scenario in the loss can enable the model to learn differences between subjects, improving subject separability. To address this limitation, we introduced the domain margin loss, adding it as the second term in Eq. (7). This modification is a critical consideration for improving subject-specific invariance by explicitly modeling inter-subject differences.

Furthermore, the cross-entropy loss was not included in the original training process, as classification was performed during fine-tuning after the FAE model had been trained solely with the reconstruction loss. However, using reconstruction error as a surrogate loss function for model hyperparameter tuning did not yield models that performed well in classification tasks²⁷. This suggests that optimizing solely for reconstruction error

is insufficient for achieving high classification performance, emphasizing the need for a more targeted approach that integrates both class separability and domain invariance from the outset.

Next to the contrastive loss, we investigated using a triplet loss²⁶ instead, which we denote as the Triplet Factorized Autoencoder (TFAE). It operates on triplets of samples: an anchor, a positive (same class as anchor) sample, and a negative (different class from anchor) sample. It learns to minimize the distance between the anchor and positive sample while maximizing the distance between the anchor and the negative sample.

The corresponding triplet class loss and domain loss are defined as:

$$L^c(z_i^d, z_j^d, z_k^d) = \max(\|z_i^c - z_j^c\|_2^2 - \|z_i^c - z_k^c\|_2^2 + m^c, 0) \quad (8)$$

$$L^d(z_i^d, z_j^d, z_k^d) = \max(\|z_i^d - z_j^d\|_2^2 - \|z_i^d - z_k^d\|_2^2 + m^d, 0) \quad (9)$$

It is important to note that sample i, j and k are the anchor, positive and negative sample, respectively. Therefore, the positive sample must have the same class and subject label as the anchor, while the negative sample must have a different class and subject label.

In the performed experiments, the factorized models are compared to a Multilayer Perceptron (MLP) baseline that is identical in number of neurons in each layer and all shared hyperparameters, to the factorized models except that its loss function includes only cross-entropy and reconstruction loss (Eqs. 4 and 5). Hyperparameter tuning using gridsearch resulted in the following optimal settings for the loss function: a value of 0.1 for α , 0.75 for β , 1.0 for γ and 0.01 for δ . For training, we used the Adam optimizer with a learning rate of 0.1, a batch size of 128 and a maximum of 100 epochs with early stopping. Finally, the optimal network architecture included a final hidden layer with 20 neurons. The MLP baseline and factorization models are based on a neural network consisting of in total 3 layers in the encoder where the number of neurons decay linearly. Thus, there are 240 neurons in the input layer, 130 neurons in the first hidden layer and 20 neurons in the second hidden layer. Furthermore, we inspect the separability of the domain and class latent space considering four scenarios: (1) a logistic regression trained on the domain latent space using the domain (subject) label, which should be able to separate subjects well in the domain space but not in the class space as that should not separate on subjects; (2) training on the domain latent space using the class label, which should not be able to separate classes both in the domain and class space as the domain space should not separate classes; (3), training on the class latent space using the label, which should be able to separate classes well in the class space but not in the domain space; (4) and training on the class latent space using the domain label, which should not be able to separate classes in both spaces. Only the latter two apply to the MLP, as it only learns one latent space.

Evaluation

To evaluate the effectiveness of the proposed methods, we employ stratified leave-10-subjects-out cross-validation and use the ROCAUC score for the classification task as the performance metric. The data is split such that no subject appears simultaneously in both the training and validation sets, preventing information leakage. Furthermore, stratification is achieved through the pairwise and triplet sampling strategies (Appendix A), which ensure that each subject is sampled an equal number of times. Additionally, the pairwise sampling guarantee that the number of pairs sampled both within subjects and between subjects is balanced. Class labels are also sampled randomly to maintain an equal representation of both classes. Similarly, the triplet sampling approach ensures a balanced number of anchors from each class while also maintaining an equal distribution of samples from each subject.

Results

Data set

Smartwatch data from the ME-TIME study (registered at clinicaltrials.gov with ID NCT05802563) was used and Table 2 show the characteristics of the development set, used to train and tune the models, as well as the test set. The diversity among subjects, is reflected in their corresponding smartwatch data. This is particularly evident in the mean and standard deviation of the heart rate, but also in the standard deviation of the (relative) heart acceleration, which exhibit substantial inter-subject variability, as shown in Fig. 2. Such variability can significantly impact the generalization performance of machine learning models.

Furthermore, we have defined several binary classification tasks, as outlined in Table 1. The night/day classification is based on the timestamps of the heart rate sensor. Windows that fall entirely between midnight and 5 AM are labeled as 'night,' while those that fall entirely between 11 AM and 6 PM are labeled as 'day.' The inactive/active classification is determined using the step counter, which records the number of steps taken per minute. Windows with an average of fewer than 5 steps per minute are classified as 'inactive,' allowing for minor step counts due to hand movement noise. Windows with more than 50 steps per minute are classified as 'active.' To maintain clear class separation, windows with step counts between 5 and 50 are excluded. The normal/abnormal SpO2 classification is based on blood oxygen saturation levels. Windows with SpO2 values below 90 are labeled 'abnormal,' while those above 95 are labeled 'normal.' Values between 90 and 95 are excluded to ensure a distinct separation between classes. Finally, the sleep stage classification is based on labels provided by the Fitbit. Only windows in which the sleep stage remains consistent throughout the entire window are considered.

Multi-subject factorization

To evaluate the models' performance on multiple subjects, the models are trained using 50 train subjects and tested on 30 test subjects. The ROCAUC for the Night versus Day classification (left) and Inactive versus Active

Characteristic		Train subjects	Test subjects
Total		50	30
Age, years	18–39	17	9
	40–54	10	5
	55–64	9	4
	> 65	14	12
Diagnosis	Reference	26	16
	Heart failure	8	5
	Persistent atrial fibrillation	3	2
	Permanent atrial fibrillation	4	1
	Paroxysmal atrial fibrillation	9	6
Sex	Male	27	16
	Female	23	14
Diabetes	18.5–24.9	20	15
	25–29.9	16	9
	> 30	14	6
Diabetes	Yes	19	5
	No	41	25
Smoking	Yes	13	4
	No	37	26
Hypertension	Yes	17	9
	No	33	21
Device	Charge 5	32	19
	Inspire 2	18	11

Table 2. Characteristics of the development and test sets.

Mean heart rate, (relative) heart acceleration and steps per subject with corresponding standard deviation

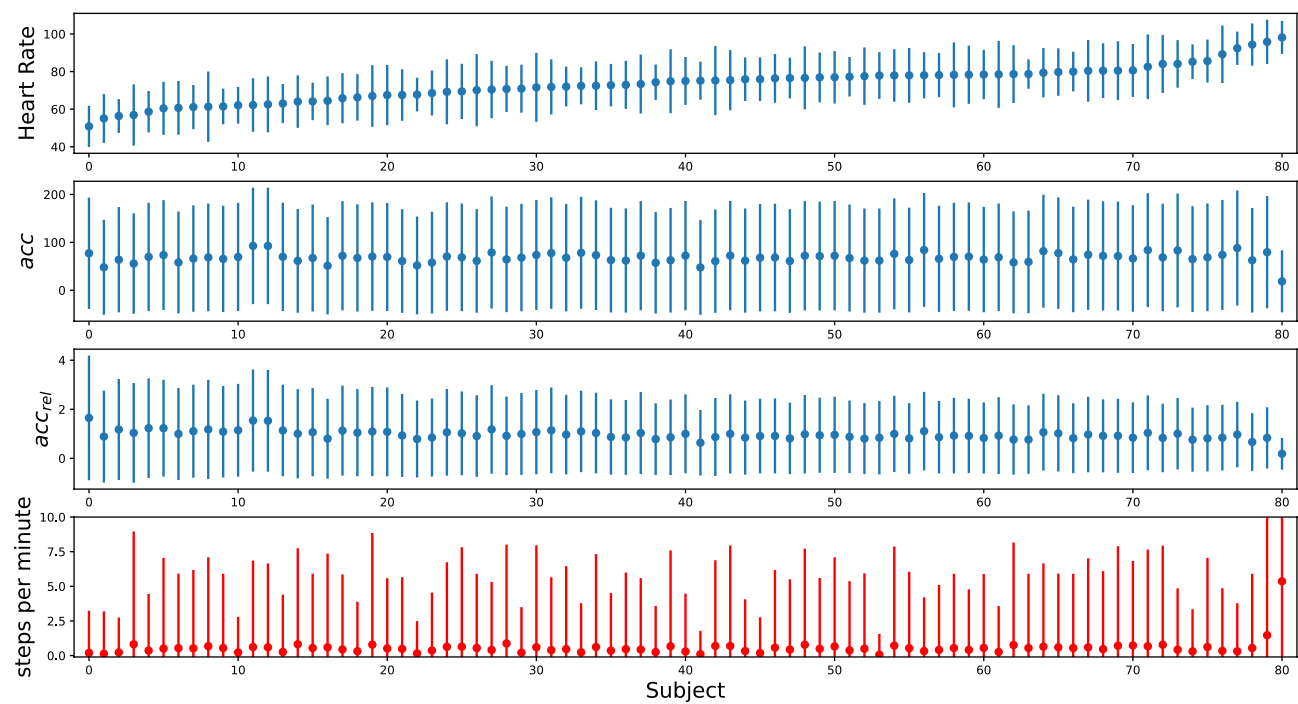


Fig. 2. Mean and standard deviation of data per subject for heart rate (first row), heart acceleration (second row), relative heart acceleration (third row) and steps per minute (last row).

Model/task	Night and day			Inactive and active		
	hr	hr,a	hr,ar	hr	hr,a	hr,ar
MLP	74.81 (10.53)	73.24 (5.83)	58.63 (7.02)	84.12 (9.41)	82.11 (7.83)	75.63 (7.22)
FAE	78.23 (9.01)	71.12 (5.34)	70.63 (5.72)	85.91 (6.41)	81.72 (8.61)	75.13 (7.82)
GFAE ($m^d = 0$)	81.24 (5.31)	73.13 (6.31)	73.23 (6.01)	87.83 (4.62)	81.61 (9.13)	74.53 (8.84)
GFAE ($m^d \neq 0$)	83.14 (5.12)	74.31 (6.12)	74.12 (5.81)	86.93 (4.42)	81.41 (9.23)	74.71 (8.63)
TFAE	83.42 (5.33)	74.21 (6.23)	74.34 (5.61)	86.62 (4.31)	81.54 (9.11)	74.73 (8.41)

Table 3. Mean and standard deviation of ROCAUC for night/day and inactive/active classification. For visual clarity, ROCAUC values are reported scaled from the standard range of 0–1 to 0–100. Hr = heart rate, hr,a = heart acceleration and hr,ar = relative heart acceleration. The GFAE is considered with a domain margin loss equal to zero ($m^d = 0$) and non-zero ($m^d \neq 0$), where after gridsearch, the optimal m^d was found to be 1. Significant values are in [bold].

Model/task	Light and restless	Light and REM	SPO2 Normal and abnormal			Light and deep sleep		
	hr	hr	hr	hr,a	hr,ar	hr	hr,a	hr,ar
MLP	61.41 (10.72)	49.63 (3.41)	51.24 (2.42)	50.31 (1.32)	49.83 (1.72)	52.93 (5.12)	48.41 (3.74)	51.72 (2.03)
FAE	62.93 (9.71)	50.02 (0.10)	50.92 (2.31)	51.71 (2.42)	50.02 (1.10)	49.01 (3.24)	49.15 (0.62)	51.82 (3.93)
GFAE ($m^d = 0$)	63.54 (8.51)	49.82 (4.73)	52.51 (3.42)	50.43 (3.71)	50.14 (2.91)	52.42 (3.31)	50.10 (1.41)	50.11 (1.32)
GFAE ($m^d \neq 0$)	63.72 (8.71)	50.14 (4.42)	52.91 (3.92)	50.03 (3.92)	50.41 (2.73)	52.23 (3.02)	50.31 (1.74)	49.93 (1.53)
TFAE	63.62 (8.42)	49.71 (4.53)	52.84 (3.72)	50.32 (3.74)	50.32 (2.51)	52.14 (3.11)	50.21 (1.64)	49.74 (1.42)

Table 4. Mean and standard deviation of ROCAUC over subjects for sleep related and SPO2 tasks. For visual clarity, ROCAUC values are reported scaled from the standard range of 0–1 to 0–100. Hr = heart rate, hr,a = heart acceleration and hr,ar = relative heart acceleration. The GFAE is considered with a domain margin loss equal to zero and non-zero. Significant values are in [bold].

classification (right) is given in Table 3. The corresponding receiver operating characteristic (ROC) curves for the unnormalized heart rate case are given in Figure E6, Appendix E.

When using the original heart rate, the factorization models show improvement on the night/day classification task and to a lesser degree, improvement on the inactive/active task. The smaller improvement gain can be accounted to the fact that the latter is an easier task, with almost 10 higher points of ROCAUC (on a 0–100 scale) for the MLP baseline.

Using heart acceleration and relative heart acceleration normalizations does not improve the models. This may be due to the fact that heart rate values at adjacent time points are frequently identical, resulting in a one-point difference of zero. Consequently, the time series becomes sparse, as illustrated in Appendix D.

Furthermore, the TFAE performs best by a margin on the night/day classification and the GFAE without domain margin loss on the inactive/active classification. However, the differences are small considering the fact that the GFAE and TFAE are within 0.5 points with a ten times higher standard deviation. Similarly, for the inactive/active classification task the GFAE and TFAE are within 1.5 points with a standard deviation of more than 4, indicating that the difference is relatively small. Overall, the standard deviation over subjects decreases when using factorization.

In Table 4, several additional tasks are considered. Both Light & REM sleep and Light & deep sleep tasks performs at near random for all models, while the normal & abnormal SPO2 performs slightly above random. The only task that performs better than random is classifying light sleep from restless sleep, where the GFAE with non-zero margin loss performs best. Furthermore, the GFAE in this configuration also has the best mean performance in three out of four tasks. Fitbit identifies restless sleep based on movement, such as tossing and turning, which notably impacts heart rate. In contrast, light and deep sleep have more subtle effects on heart rate that the Fitbit fails to capture²⁸.

Per-subject calibration

We investigated the effects of the normalization methods described in “Normalization” section. The results for Z-normalization applied per test subject, using varying percentages of each test subject’s calibration set (yellow part of Fig. 1b), are presented in Fig. 3. Figure 4 illustrates the results of Z-normalization applied per train subject, where train subjects were normalized individually, in combination with per-test-subject normalization.

Per-test subject normalization does not consistently improve performance. This is most pronounced in the GFAE with domain margin loss and the TFAE without inactivity normalization. Furthermore, normalization with too little calibration data can have adverse effects as can be seen by the dip in performance with a calibration

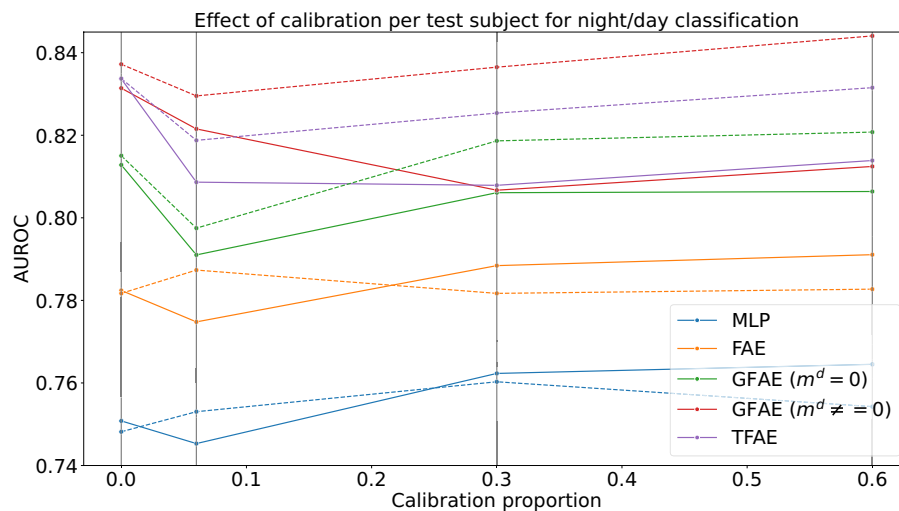


Fig. 3. Effect of calibration per-test-subject for night/day classification. Calibration proportion is the amount of data from a test subject used to compute a subject-specific mean and variance, to standardize the remaining data. Vertical grey lines denote the tested calibration proportions: 0%, 6%, 30%, and 60%. Dashed lines denote inactivity normalization.

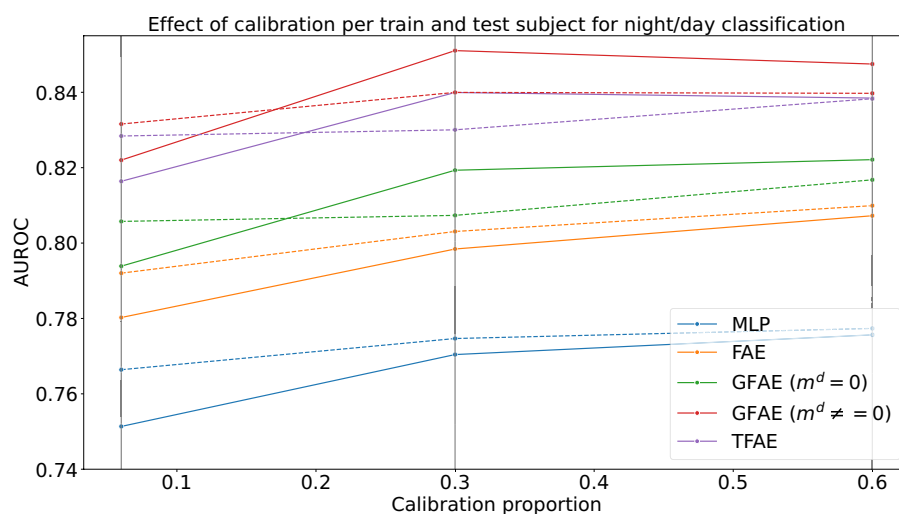


Fig. 4. Effect of calibration per-train and per-test-subject for night/day classification. Calibration proportion is the amount of data from a train or test subject used to compute a subject-specific mean and variance, to standardize the remaining data. Vertical grey lines denote the tested calibration proportions: 0%, 6%, 30%, and 60%. Dashed lines denote inactivity normalization. Since a population mean and standard deviation are not applicable in this configuration, it can only be performed with a non-zero calibration proportion.

proportion of 10%. Using all calibration data seems to benefit the mean performance for most models. In contrast, applying per-train subject normalization in addition to per-subject normalization provides a more consistent increase in performance with increasing calibration proportion.

Both configurations have been examined with and without the inclusion of inactivity normalization, which improves all models, albeit not consistently over all calibration proportions. The GFAE is the best performing model. When calibration is done per test subject, inactivity normalization is required to reach a similar performance as calibration per train and test subject.

Class and domain latent space analysis

To quantify how well class and domain information is disentangled into their corresponding latent space, a logistic regression was trained on both latent spaces using either class labels, to inspect how well the classes are separated (class accuracy) or domain labels, to inspect how well the subjects can be separated (domain accuracy). Ideally, the class latent space should excel at class label prediction but perform poorly on subject labels, while the domain latent space should excel at subject label prediction but perform poorly on class labels.

Model	Latent space	Class accuracy	Domain accuracy
MLP	z^c	89.71	8.52
FAE	z^c	82.42	10.21
	z^d	74.14	10.93
GFAE ($m^d = 0$)	z^c	90.52	14.11
	z^d	87.70	16.24
GFAE ($m^d \neq 0$)	z^c	98.61	25.62
	z^d	97.32	36.94
TFAE	z^c	98.43	10.23
	z^d	98.21	28.14

Table 5. Night/day class and domain train ROCAUC of logistic regression on z^c and z^d using 50 train subjects.

Model	Latent space	Class accuracy	Domain accuracy
MLP	z^c	76.81	9.12
FAE	z^c	80.23	3.81
	z^d	76.12	4.53
GFAE ($m^d = 0$)	z^c	82.21	10.24
	z^d	76.71	11.01
GFAE ($m^d \neq 0$)	z^c	83.02	12.22
	z^d	73.10	14.70
TFAE	z^c	84.44	17.74
	z^d	74.31	20.11

Table 6. Night/day class and domain test ROCAUC of logistic regression on z^c and z^d using 30 test subjects.

The train and test results for night day classification are presented in Tables 5 and 6. Similarly, the results for inactive active classification can be found in Appendix C, tables C1 and C2.

Interestingly, for the MLP model, the domain accuracy is slightly lower on the training set than on the test set, likely due to the increased complexity of the training set classification task. Nevertheless, all models, including the baseline MLP, surpass the 2% and 3.3% baseline accuracy for the train and test performance respectively (as a results of having 50 subjects in the train set and 30 subjects in the test set), demonstrating their ability to extract meaningful patterns despite the challenges posed by the subject distribution.

The results indicate further that both latent spaces contain substantial class information, as evidenced by their strong performance on the class labels. This can also be observed in the first and third column of the Uniform Manifold Approximation and Projection (UMAP) (Fig. 5, C3 and C4), where the class and subject latent spaces are colored by class label. Still a significant class difference can be observed in z^d . Similarly, the domain latent space of factorization models often outperform the MLP in subject classification, suggesting a correlation between domain and class features, which could stem from the fact that the loss functions of the factorization models do not explicitly enforce feature decorrelation. although the factorization model tries to explicitly put the subject information on the domain latent space. Further analyses on pairwise subject classification (Appendix B), confirm this observation.

For both the train and test sets, the GFAE with positive domain margin loss as well as the TFAE improve domain accuracy using z^d compared to the MLP. This demonstrates that the factorization models effectively capture and encode subject-specific information in the domain latent space. However, these improvements are less pronounced in the test set, particularly for the inactive/active task, suggesting that the factorization models could benefit from more subjects to generalize better.

Discussion and conclusion

We have investigated the challenge of inter-subject variability in heart rate time series, that limits the generalizability of machine learning models to unseen subjects. By evaluating normalization techniques and factorization models, we aimed to reduce this variability and improve classification performance for tasks such as night/day and inactive/active classification.

The heart rate time series that were transformed by the (relative) first-difference showed degraded performance, likely due to the sparsity it introduces in the signal, as heart rate values at adjacent time points are frequently identical, resulting in zeros or a one-point difference of zero. Factorized autoencoder models demonstrated consistent improvements, particularly for night/day classification. The GFAE with a positive

Zclass and Zdomain embedding of night/day classification

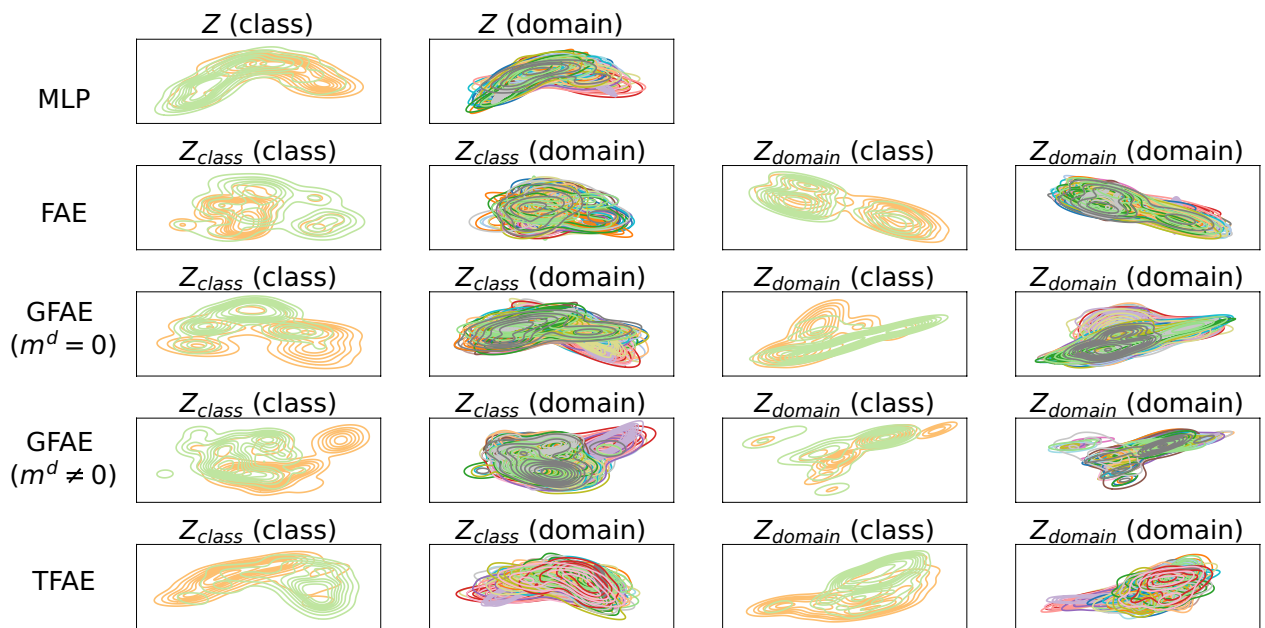


Fig. 5. UMAP visualization of the models using the test data for night/day classification. Parentheses indicate coloring (class or domain).

domain margin loss as well as the TFAE achieved the highest accuracy, highlighting the benefit of explicitly modeling inter-subject variability in the loss function.

Using the factorization models as base, we calibrated them for each subject individually, using either Z-normalization on all the heart rate data or only during inactive moments. We investigated this calibration in two situations: calibrating only test subjects using a withheld calibration set or calibrating both train and test subjects. Combining normalization by the heart rate during inactive periods (referred to as inactivity normalization) with calibration across both train and test subjects yielded the most consistent improvements.

The factorization models were further investigated by using a logistic regression to investigate the separability of the classes in the class-related latent space and the subjects in the subject-related latent space.

The factorized models improved separability of subjects in the subject-related space. However, while the improvements over the baseline were significant, the separability still leaves substantial room for further refinement. Latent space analysis revealed that class-related information was encoded in both task and domain-latent spaces, potentially due to the loss functions' lack of explicit constraints to enforce feature decorrelation. Future work could explore techniques to enforce this by incorporating an adversarial component, as demonstrated in the sensor-based human activity recognition literature^{29,30}, could be beneficial.

The addition of a domain margin loss allowed the models to learn differences between subjects, which is critical for addressing inter-subject variability. This loss explicitly models the distinctions between subjects in the subject-related latent space, enabling the model to better disentangle subject-specific characteristics. Future work could extend this by leveraging metadata on subjects, such as demographic and physiological information, directly in the domain margin loss.

Unfortunately, tasks such as sleep state and SpO2 classification, showed poor performance, where only the restless/light sleep task performed slightly better than random. The limited utility of these tasks is likely due to the lack of reliable labels from consumer-grade devices like Fitbit³¹. Furthermore, sleep stage is estimated using body movement data from the accelerometer and heart rate variability (HRV) derived from the PPG sensor—granular information that may not be fully preserved in the estimated heart rate in smartwatches.

By advancing methods to handle inter-subject variability, as demonstrated in this study, machine learning applications in smartwatch health monitoring will be better equipped to generalize across diverse populations. Our proposed generalized factorized autoencoder and triplet factorized autoencoder showed improvements using smartwatch data, highlighting their potential to address inter-subject variability. These advancements not only contribute to more accurate and reliable models but also pave the way for personalized and inclusive health monitoring solutions, ensuring greater applicability to real-world scenarios.

Data availability

The anonymized heart rate, step counter and deidentified timestamps of the Fitbit time series data are available on request to a.naserijahfari@hagaziekenhuis.nl with a signed data access agreement.

Received: 28 February 2025; Accepted: 15 July 2025

Published online: 23 July 2025

References

- Sabry, F., Eltaras, T., Labda, W., Alzoubi, K. & Malluhi, Q. Machine learning for healthcare wearable devices: The big picture. *J. Healthc. Eng.* **2022**(1), 4653923 (2022).
- Saad, H. S., Zaki, J. F. & Abdelsalam, M. M. Employing of machine learning and wearable devices in healthcare system: Tasks and challenges. *Neural Comput. Appl.* **36**(29), 17829–17849 (2024).
- Noitz, M. et al. Detection of subtle ECG changes despite superimposed artifacts by different machine learning algorithms. *Algorithms* **17**(8), 360 (2024).
- World Health Organization: Cardiovascular diseases (CVDs). Accessed: 2025-02-04 (2021). <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-cvds>
- Townsend, N. et al. Epidemiology of cardiovascular disease in Europe. *Nat. Rev. Cardiol.* **19**(2), 133–143 (2022).
- McDonagh, T. A. et al. European society of cardiology heart failure association standards for delivering heart failure care. *Eur. J. Heart Fail.* **13**(3), 235–241 (2011).
- Mayo Clinic Staff: Heart Disease: Diagnosis and Treatment. Accessed: 2025-02-04 (2024). <https://www.mayoclinic.org/diseases-conditions/heart-disease/diagnosis-treatment/drc-20353124>
- Lauer, M. S., Anderson, K. M., Kannel, W. B. & Levy, D. The Impact of Obesity on Left Ventricular Mass and Geometry: The Framingham Heart Study. *JAMA* **266**(2), 231–236. <https://doi.org/10.1001/jama.1991.03470020057032> (1991).
- Ryan, S. M., Goldberger, A. L., Pincus, S. M., Mietus, J. & Lipsitz, L. A. Gender- and age-related differences in heart rate dynamics: Are women more complex than men?. *J. Am. Coll. Cardiol.* **24**(7), 1700–1707. [https://doi.org/10.1016/0735-1097\(94\)90177-5](https://doi.org/10.1016/0735-1097(94)90177-5) (1994).
- Oberman, A., Myers, A. R., Karunas, T. M. & Epstein, F. H. Heart size of adults in a natural population-tecumseh, michigan: Variation by sex, age, height, and weight. *Circulation* **35**(4), 724–733 (1967).
- Liao, D. et al. The ARIC Investigators: Age, race, and sex differences in autonomic cardiac function measured by spectral analysis of heart rate variability-the aric study. *Am. J. Cardiol.* **76**(12), 906–912. [https://doi.org/10.1016/S0002-9149\(99\)80260-4](https://doi.org/10.1016/S0002-9149(99)80260-4) (1995).
- Quer, G., Gouda, P., Galarnyk, M., Topol, E. J. & Steinhubl, S. R. Inter- and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, BMI, and time of year: Retrospective, longitudinal cohort study of 92,457 adults. *PLoS ONE* **15**(2), 0227709 (2020).
- Chen, Y. et al. Smartwatch-based algorithm for early detection of pulmonary infection: Validation and performance evaluation. *Digit. Health* **10**, 20552076241290684 (2024).
- Tison, G. H. et al. Passive detection of atrial fibrillation using a commercially available smartwatch. *JAMA Cardiol.* **3**(5), 409–416 (2018).
- Naseri, A., Tax, D. M., Reinders, M. & Bilt, I. Heart disease detection using an acceleration-deceleration curve-based neural network with consumer-grade smartwatch data. *Heliyon* **10**(21), e39927 (2024).
- Wasserlauf, J. et al. Smartwatch performance for the detection and quantification of atrial fibrillation. *Circ. Arrhythm. Electrophysiol.* **12**(6), 006834 (2019).
- Zhou, Z., Zhang, Y., Yu, X., Yang, P., Li, X.-Y., Zhao, J. & Zhou, H. Xhar: Deep domain adaptation for human activity recognition with smart devices. In: *2020 17th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pp. 1–9 (2020). IEEE
- Zhu, T., Afentakis, I., Li, K., Armiger, R., Hill, N., Oliver, N. & Georgiou, P. Multi-horizon glucose prediction across populations with deep domain generalization. *IEEE J. Biomed. Health Inform.* (2024)
- Shen, X. et al. Multi-omics microsampling for the profiling of lifestyle-associated changes in health. *Nat. Biomed. Eng.* **8**(1), 11–29 (2024).
- Kumar, C. S., Ramachandran, K. & Kumar, A. Vital sign normalisation for improving performance of multi-parameter patient monitors. *Electron. Lett.* **51**(25), 2089–2090 (2015).
- Gyawali, P. K., Horacek, B. M., Sapp, J. L. & Wang, L. Sequential factorized autoencoder for localizing the origin of ventricular activation from 12-lead electrocardiograms. *IEEE Trans. Biomed. Eng.* **67**(5), 1505–1516 (2019).
- Cai, R., Li, Z., Wei, P., Qiao, J., Zhang, K. & Hao, Z. Learning disentangled semantic representation for domain adaptation. In: *IJCAI: Proceedings of the Conference*, vol. 2019, p. 2060 (2019). NIH Public Access
- Ilse, M., Tomczak, J. M., Louizos, C. & Welling, M. Diva: Domain invariant variational autoencoders. In: *Medical Imaging with Deep Learning*, pp. 322–348 (2020). PMLR
- Röttgering, J. G. et al. Determining a target spo2 to maintain pao2 within a physiological range. *PLoS ONE* **16**(5), 0250740 (2021).
- Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In: *International Conference on Machine Learning*, pp. 1597–1607 (2020). PMLR
- Hoffer, E. & Ailon, N. Deep metric learning using triplet network. In: *Similarity-based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12–14, 2015. Proceedings 3*, pp. 84–92 (2015). Springer
- Kouw, W. M. & Loog, M. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806* (2018)
- Tsunoda, M., Endo, T., Hashimoto, S., Honma, S. & Honma, K.-I. Effects of light and sleep stages on heart rate variability in humans. *Psychiatry Clin. Neurosci.* **55**(3), 285–286 (2001).
- Su, J., Wen, Z., Lin, T. & Guan, Y. Learning disentangled behaviour patterns for wearable-based human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **6**(1), 1–19 (2022).
- Qian, H., Pan, S. J. & Miao, C. Latent independent excitation for generalizable sensor-based cross-person activity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence* **35**, 11921–11929 (2021).
- Haghighyegh, S., Khoshnevis, S., Smolensky, M. H., Diller, K. R. & Castriotta, R. J. Accuracy of wristband fitbit models in assessing sleep: systematic review and meta-analysis. *J. Med. Internet Res.* **21**(11), 16273 (2019).

Author contributions

A.N., D.T. and M.R. conceptualized the study and developed the methodology. Data curation, formal analysis and software development was performed by A.N. Supervision was done by D.T, M.R and I.B. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-12102-7>.

Correspondence and requests for materials should be addressed to A.N.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/>.

© The Author(s) 2025