Self-supervised transfer learning of physiological representations from free-living wearable data

Dimitris Spathis
University of Cambridge, UK
Department of Computer Science and Technology
ds806@cl.cam.ac.uk

Ignacio Perez-Pozuelo*
Soren Brage
Nicholas J. Wareham
University of Cambridge, UK
MRC Epidemiology Unit,
School of Clinical Medicine
{ip325,sb400,njw1004}@cam.ac.uk

Cecilia Mascolo
University of Cambridge, UK
Department of Computer Science and Technology
cm542@cl.cam.ac.uk

ABSTRACT

Wearable devices such as smartwatches are becoming increasingly popular tools for objectively monitoring physical activity in free-living conditions. To date, research has primarily focused on the purely supervised task of human activity recognition, demonstrating limited success in inferring high-level health outcomes from low-level signals. Here, we present a novel *self-supervised* representation learning method using activity and heart rate (HR) signals without semantic labels. With a deep neural network, we set HR responses as the *supervisory signal* for the activity data, leveraging their underlying physiological relationship. In addition, we propose a custom quantile loss function that accounts for the long-tailed HR distribution present in the general population.

We evaluate our model in the largest free-living combined-sensing dataset (comprising >280k hours of wrist accelerometer & wearable ECG data). Our contributions are two-fold: i) the pre-training task creates a model that can accurately forecast HR based only on cheap activity sensors, and ii) we leverage the information captured through this task by proposing a simple method to aggregate the learnt latent representations (embeddings) from the window-level to user-level. Notably, we show that the embeddings can generalize in various downstream tasks through transfer learning with linear classifiers, capturing physiologically meaningful, personalized information. For instance, they can be used to predict variables associated with individuals' health, fitness and demographic characteristics (AUC >70), outperforming unsupervised autoencoders and common bio-markers. Overall, we propose the first multimodal self-supervised method for behavioral and physiological data with implications for large-scale health and lifestyle monitoring.
Code: `https://github.com/sdimi/Step2heart`.

*Also with the Alan Turing Institute, UK

ACM CHIL '21, April 8–10, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8359-2/21/04.
https://doi.org/10.1145/3450439.3451863

CCS CONCEPTS

• **Computing methodologies** → **Unsupervised learning**; *Transfer learning*; • **Mathematics of computing** → *Time series analysis*; • **Human-centered computing** → *Ubiquitous and mobile devices.*

ACM Reference Format:
Dimitris Spathis, **Ignacio Perez-Pozuelo**, **Soren Brage**, **Nicholas J. Wareham**, and **Cecilia Mascolo**. 2021. Self-supervised transfer learning of physiological representations from free-living wearable data . In *ACM Conference on Health, Inference, and Learning (ACM CHIL '21), April 8–10, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3450439.3451863

1 INTRODUCTION

The advent of wearable technologies has given individuals the opportunity to unobtrusively track everyday behavior. Given the rapid growth in adoption of internet-enabled wearable devices, sensor time-series comprise a considerable amount of user-generated data [5]. However, extracting meaning from this data can be challenging, since sensors measure low-level signals (e.g., acceleration) as opposed to the more high-level events that are usually of interest (e.g., arrhythmia, infection or obesity onset). Most wearable devices, particularly those that are wrist-worn, incorporate accelerometry sensors, which are very affordable tools to objectively study physical activity patterns [15, 37][1]. However, since wearables are used in daily, unconstrained environments, activities like drinking coffee or alcohol, as well as stress, may confound simple heuristics.

Deep learning models, on the other hand, do not require handcrafted feature engineering, capture the temporal dynamics of sequential data, and exploit the latent representations inherently present in this data [20]. These approaches have shown great promise in human activity recognition (HAR) tasks using wearable sensor data [2, 31, 59], but rely on purely labeled datasets which are costly to collect [7]. In addition, they are obtained in laboratory settings and hence might not generalize to free-living conditions where behaviours are more diverse, covering a long tail of activities [27].

Unsupervised learning is a natural candidate to solve this label scarcity problem in wearable data, particularly given the vast amounts that can be collected in free-living conditions. Recent models have effectively utilized unlabeled activity data to learn useful summary representations of sensor signals [1]. Notwithstanding

[1]Throughout this work, we refer to activity, movement and acceleration interchangeably as signals obtained from wearable accelerometers.

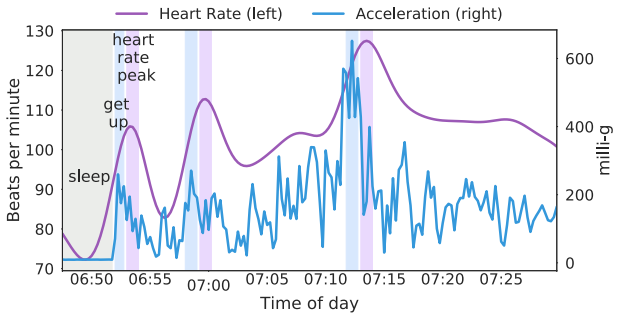

Figure 1: Heart rate and acceleration temporal dynamics. Illustrative visualization of the relationship between movement and heart rate responses (randomly selected participant). Shaded areas show this lagging relationship.

the value of these newly proposed methods, they only rely on a single stream of sensor data, usually movement data, and do not fully exploit the multimodal nature of modern wearable devices.

Indeed, physical activity is characterized by *both* movement and the associated cardiovascular response to movement (e.g., heart rate increases after exercise and the dynamics of this increase are dictated by fitness levels [24]), thus, leveraging these two signals concurrently likely produces better representations than either signal taken in isolation. This relationship is conceptualized in Figure 1. Heart rate (HR) responses to exercise have been shown to be strongly predictive of cardiovascular disease (CVD), coronary heart disease (CHD) and all-cause mortality [48]. In healthy individuals, HR responses to activity are defined by an increase in HR that is concurrent to the increasing intensity of the activity [16].

Multimodal learning has proven beneficial in supervised tasks such as fusing images with text to improve word embeddings [35], video with audio for speech classification [38], or different sensor signals for HAR [43]. However, all these approaches rely on the modalities being used as parallel inputs, limiting the scope of the resulting representations. Self-supervised training allows for mappings of aligned coupled data streams (e.g. audio to images [41] or, in our case, activity to heart rate), using unlabeled data with supervised objectives [28].

In this work, we present *Step2Heart*, a general-purpose self-supervised feature extractor for wearable data, which leverages the multimodal nature of modern wearable devices to generate participant-specific representations. This architecture can be broken into two parts: 1) The new pre-training task forecasts ECG-level quality HR in real-time by only utilizing activity signals, 2) then, we leverage the learned representations of this model to predict personalized health-related outcomes through transfer learning with *linear* classifiers. We hypothesize that this mapping captures more meaningful information than autoencoders trained on activity data or other bio-markers.

This paper puts forward four key technical contributions:

- We propose a novel *self-supervised* model and a pre-training task which maps activity data to HR responses. Through this architecture, our model learns *physiologically meaningful* user-level representations that can then be used for a variety of practical downstream tasks that are *personalized* to the users' unique physiology.
- For pre-training, we introduce a joint *loss function* that acts as a regularizer to traditional MSE by using the quantiles of the predictive density of the model in order to approximate the long-tails of HR data, an ubiquitous problem in real-world (health) data.
- We evaluate this model in the largest multimodal wearable ECG and wrist accelerometry dataset, including over 1,700 participants tracked for a week, along with associated health outcomes measured with clinical lab equipment. We perform ablation tests to show the performance of different modalities and components to the architecture.
- We perform a set of downstream, transfer learning tasks by aggregating the window-level features to user-level ones and showcase the value captured by the learned *embeddings* through strong performance at inferring physiologically meaningful variables, outperforming autoencoders and common bio-markers. For example, our models achieve an AUC of 0.70 for Body Mass Index (BMI) prediction and an AUC of 0.80 for Physical Activity Energy Expenditure.

We envision our work having applications in facilitating the comprehensive monitoring of cardiovascular health and fitness at scale. Further, our models could be used to correct faulty HR readings of noisy sensors such as PPGs and broadly to characterize the objectively measured physical behaviours in large population cohorts. Some of the downstream classification tasks highlight the potential of these techniques for the monitoring of important health information, which is usually costly or burdensome to obtain (such as fitness or obesity levels). The proposed model is summarized in Figure 2 and our code/models are available here `https://github.com/sdimi/Step2heart`.

2 RELATED WORK

Objective monitoring of physical behaviors. Large scale studies of physical activity leveraging mobile devices' built-in accelerometers have shown promise as global physical activity surveillance tools, demonstrating inequality across different countries and world regions [3]. Mobile and wearable sensors allow for continuous and ubiquitous monitoring of an individual's physical activity profiles, which combined with cardio-respiratory information, provides valuable insights into that individuals' health and fitness status [33]. Hence, the possibility of measuring individuals' physiological characteristics in free-living conditions is of great interest for research, clinical and commercial applications.

Machine learning for wearable sensing. Recently, advances in deep learning architectures for sequential modeling based upon wearable and mobile sensing have been used for health predictions and recommendations [4, 49, 52]. For example, *FitRec*, an LSTM-based approach to modelling HR and activity data for personalized fitness recommendations was able to learn activity-specific contextual and personalized dynamics of individual user HR profiles during exercise segments [39]. This approach is helpful but requires prior segmentation of activities, which can be a constraint when applying these techniques in free-living, unconstrained conditions.

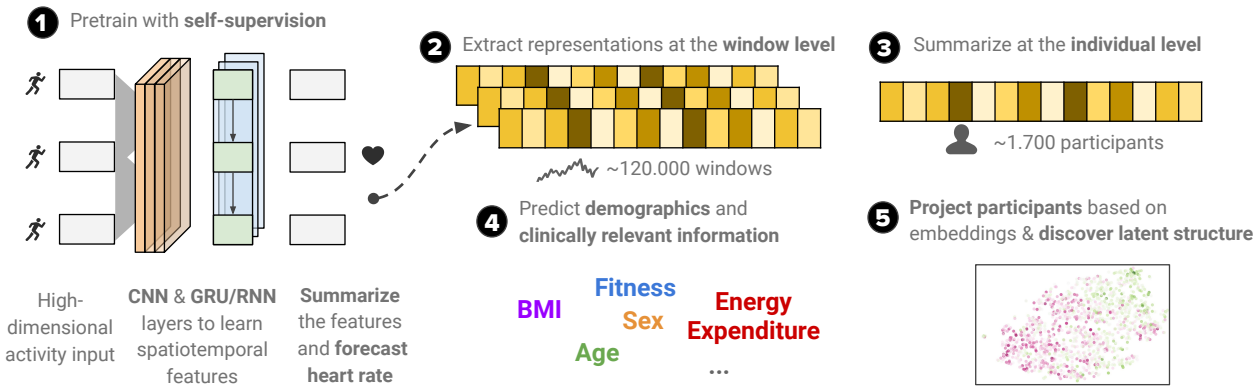

Figure 2: Schematic of model architecture and tasks.

Additionally, previous work has explored forecasting heart rate from movement data, however this was done on a much smaller scale (3 users) and used PPG sensors instead of the more accurate ECG as ground-truth [36].

Self-supervised pre-training. Recent work using self-supervised learning has shown state-of-the-art results in computer vision [23, 30], signal processing and natural language processing [28]. Use cases in wearable and mobile sensing have been limited to human activity recognition using mobile devices [45, 56] and emotion recognition using ECG data [47], both using a single modality (acceleration or ECG), whereas we explore the unsupervised combination thereof guided by their physiological relationship. Our work is also inspired by the cardiovascular signature network introduced by Hallgrimmson et al [19]. However, this is an auto-encoder based approach requiring a historical input of 1-month of data for its prediction which renders the whole setup not feasible for real time applications. Furthermore, the data used is much more aggregated and limited in terms of outcomes than the data presented here. Overall, the generalizability of the learned embeddings is an under-explored area with some recent promising results in hospital operation room data [9], while abstract (non-sensor related) attributes like gender and age have been proved to be predictable with wearable embeddings [58]. In sum, we believe that our work is the first multimodal general-purpose model to extract physiological and behavioral representations[2].

3 METHODS

In this section, we provide a brief introduction to the problem formulation and notation used and then explore the model architecture and the associated methods proposed in this work.

Problem formulation and notation. For this work, we assume N samples of T timesteps and F features of an input dataset $\mathbf{X} = (\mathbf{x}_1,...,\mathbf{x}_N) \in \mathbb{R}^{N\times T\times F}$ and a target heart rate response $\mathbf{y} = (\mathbf{y}_1,...,\mathbf{y}_N) \in \mathbb{R}^N$. Additionally, we also consider contextual metadata like the hour of the day $\mathbf{M} = (\mathbf{m}_1,...,\mathbf{m}_N) \in \mathbb{R}^{N\times F}$. We use the same length T for all sequences in our model. However, this sequence length is not a requirement and can be adapted based on the requirements of the task at hand or the granularity of the data. The intermediate representations of the model after training are $\mathbf{E} = (\mathbf{e}_1,...,\mathbf{e}_N) \in \mathbb{R}^{N\times D}$ where D is the latent dimension. These embeddings are aggregated at the user level $\tilde{\mathbf{E}} = (\tilde{\mathbf{e}_1},...,\tilde{\mathbf{e}_N}) \in \mathbb{R}^{\frac{N}{U}\times D}$, where U is the number of users, in order to predict relevant outcome variables $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}_1},...,\tilde{\mathbf{y}_N}) \in \mathbb{R}^N$. Our full notation is summarized in Table 1. We employ two representation learning tasks: self-supervised pre-training and a downstream transfer learning task.

Upstream task: self-supervised pre-training and HR forecasting. Given the accelerometer input sensor sequence $\mathbf{X}$ and associated metadata $\mathbf{M}$, predict the target HR $\mathbf{y}$ in the future. The input and target data shouldn't share temporal overlap in order to leverage the cardiovascular responses with the self-supervised paradigm by learning to predict the future. Similar formulations have been proposed in mental health forecasting [53] and reinforcement learning for video prediction [18]. Motivated by population differences in heart rates, here we propose a custom *quantile regression loss* to account for the tails of the distribution. This task by itself can be used for a reliable and real-time estimation of HR based on activity data.

Downstream task: transfer learning of learned physiological representations. Given the internal representations $\mathbf{E}$ –usually at the penultimate layer of the aforementioned neural network [46]–, predict relevant variables $\tilde{\mathbf{y}}$ regarding the users' fitness and health using traditional classifiers (e.g. Logistic Regression). Inspired by the associations between word and document vectors in NLP [29], we develop a simple aggregation method of sensor windows to the user level. This is a common issue in the literature [10].

3.1 Model architecture

As shown in Figure 2 we propose *Step2Heart*, a deep neural network for HR forecasting and transfer learning. Its layers receive high-dimensional activity inputs along with associated metadata and learn spatio-temporal dynamics in order to accurately predict HR responses. It uses stacked convolutional (CNN) and recurrent (RNN) layers building upon architectures like *DeepSense* [60], which have been proven state of art in mobile sensing. Here we present each component of the model. An overview of the overall method is given as a pseudocode in Algorithm 1. We note that we do not

[2]We note that some results appeared as extended abstract in the ML4MH workshop at NeurIPS 2020 [51].

Notation	Description
$\mathscr{D}_{train}$, $\mathscr{D}_{test}$	training and testing set for the forecasting task
$\mathbf{X}$, $\in \mathbb{R}^{N \times T \times F}$	input sensor sequences
$\mathbf{M}$, $\in \mathbb{R}^{N \times F}$	input user metadata
$\mathbf{y}$, $\in \mathbb{R}^{N}$	target heart rate response
N	number of data points (samples)
T	length of input sequence
F	number of features (attributes)
U	number of users
$\tilde{\mathscr{D}}_{train}$, $\tilde{\mathscr{D}}_{test}$	training and testing set for the transfer learning task
θ	parameters (weights) of a trained neural network
D	dimension of latent space embedding
$\mathbf{E}$, $\in \mathbb{R}^{N \times D}$	embeddings matrix learned from activity to heart rate mapping
$\tilde{\mathbf{E}}$, $\in \mathbb{R}^{U \times D}$	embeddings matrix learned like $\mathbf{E}$ (aggregated at the user level)
$\tilde{\mathbf{y}}$, $\in \mathbb{R}^{U}$	target variable for transfer learning (user level)

Table 1: Notation.

claim novelty on the backbone model and its layers, instead, we keep its architecture as simple as possible in order to showcase that the task of mapping activity to (future) heart rate signals with a joint quantile loss enables the model to learn generalizable representations of the users' current health state, which can generalize in different downstream tasks.

3.1.1 CNNs to learn spatial features. Given an input dataset $\mathbf{X}$ = ($\mathbf{x}_1$,...,$\mathbf{x}_N$), it passes through a stack of CNN layers that scan over the sequences with 1D windows and learn filters $f : \{0, ..., k-1\} \in \mathbb{R}$. The convolution operation C of a sequence element s is defined as

$$C(s) = (\mathbf{x} * f)(s) = \sum_{i=0}^{k-1} f(i) \cdot \mathbf{x}_{s-i} \tag{1}$$

where k is the filter size, $s - i$ records the convolution step and $*$ denotes the convolution operator. Please note that the 1D window learns patterns across all the parallel features of the 3D input tensor $\mathbf{X}$.

3.1.2 RNNs to learn temporal features. The learned filters of the CNNs are then fed into stacked RNNs. Specifically we employ a fast variant of RNNs known as Gated Recurrent Units (GRU) [13]. The GRU has a reset gate r and an update gate z which change the hidden state h at each time step. The update functions are as follows:

$$\begin{aligned}
\mathbf{r}_t &= \sigma(W_r \mathbf{x}_t + U_r \mathbf{h}_{t-1} + \mathbf{b}_r) \\
\mathbf{z}_t &= \sigma(W_z \mathbf{x}_t + U_z \mathbf{h}_{t-1} + \mathbf{b}_z) \\
\tilde{\mathbf{h}_t} &= tanh(W \mathbf{x}_t + U(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}) \\
\mathbf{h}_t &= (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}_t}
\end{aligned} \tag{2}$$

where matrices W, U and $\mathbf{b}$ are model parameters and biases respectively, σ is a sigmoid function, and $\odot$ element-wise multiplication. The stacked GRUs output sequences which correspond to latent temporal features.

Algorithm 1: *Step2Heart* model pseudocode

```
Input  : X (sensors), M (metadata), y (target HR)
Output : Ẽ (user-level embedding), ỹ (target variable)
while neural network θ not converged do
    pass X through CNN/RNN layers (eq. 1 & 2);
    pass M through reLU layers;
    concatenate outputs in E;
    forecast & backpropagate with joint loss ℒ (eq. 5);
end
use trained network θ to extract embeddings E;
aggregate E to the user-level Ẽ with average pooling;
train a linear model to predict target variables ỹ;
```

3.1.3 Pooling and prediction. Then, the GRU output $\mathbf{h}_t$ passes through a pooling layer that performs global element-wise averaging in order to summarize all the timesteps of the 3D tensor to a 2D matrix. If needed, the representation after the pooling operation can be concatenated with other features or metadata after passing through feed forward *ReLu* layers. We also refer to this representation at the penultimate layer, $\mathbf{E}$, or *embeddings* matrix. Lastly, the final layer is a feed forward neural network with a linear activation which is appropriate for regression tasks.

3.2 Loss function

Heart rates vary across large populations. As such, some individuals may reach very low (<50 bpm, at rest/sleeping) or high (>180 bpm during vigorous exercise) [55] generating very long tails on the heart rate distribution. In traditional regression, the aim is to minimize the squared-error loss function or MSE $\mathcal{L}_{MSE}(\mathbf{y}, \mathbf{f}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(y_i - f(\mathbf{x}_i))^2$ to predict a single point estimate, similarly, quantile regressions aim to minimize the quantile loss in predicting a certain quantile. As such, the higher the quantile, the more the quantile loss function penalizes underestimates and the less it penalizes over estimates.

The loss for an individual data point in quantile regression is defined by:

$$\mathcal{L}(\xi_i|\alpha) = \begin{cases} \alpha\xi_i & \text{if } \xi_i \geq 0, \\ (\alpha - 1)\xi_i & \text{if } \xi_i < 0. \end{cases} \tag{3}$$

where α is the required quantile (between 0 and 1) and $\xi_i = y_i - f(x_i)$, where $f(x)$ is the predicted (quantile) model and y is defined by the observed value for input x. A more compact version of Eq. (3) can be formulated as $\mathcal{L}(\xi_i|\alpha) = max(\alpha\xi_i, (\alpha-1)\xi_i)$ where $\xi \in \mathbb{R}$ is the residual. As such, the average quantile loss over the whole dataset is:

$$\mathcal{L}_Q(\mathbf{y}, \mathbf{f}|\alpha) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(y_i - f(\mathbf{x}_i)|\alpha) \tag{4}$$

The quantile loss (or tilted/pinball loss in the literature) can be seen as *tilted* version of the l_1 loss which estimates the unconditional median. Instead, if a prediction falls below a given quantile

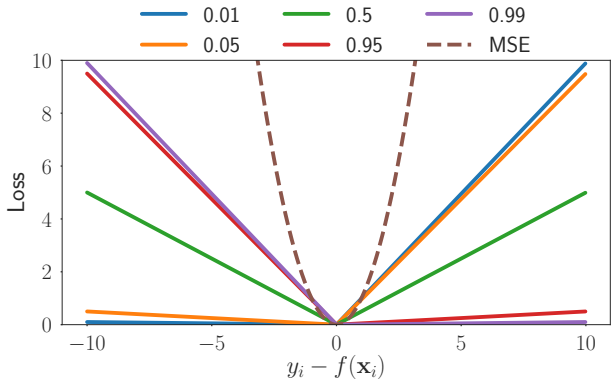

Figure 3: Quantile vs MSE loss. Illustration of the relationship between the prediction and the loss with respect to the shapes of the MSE and various levels α of quantiles. Simulated data, the true value is $y_i = 0$.

(e.g. $\alpha = 0.10$), the residual is scaled (or tilted) by its probability α. Thus, we can obtain the conditional quantile by minimizing the empirical $\mathcal{L}_Q$ loss. This formulation is inspired by similar loss functions applied to transportation problems [44] as well as reinforcement learning [14].

In practice, we are interested in various quantile levels for the predicted probability distribution, not only one. Let $\{a\}_{j=1}^{J}$ be a set of J quantiles (e.g. 0.05, 0.10, ..) we propose a joint loss function that leverages the $\mathcal{L}_{MSE}$ and $\mathcal{L}_Q$ loss for an arbitrary number of quantiles:

$$\mathcal{L}_{MSE+Q} = \frac{1}{N}\sum_{i=1}^{N}\Bigg((y_i - f(\mathbf{x}_i))^2 + \sum_{j=1}^{J} max\Big(\alpha_j(y_i - f(\mathbf{x}_i)^{(\alpha_j)}), (\alpha_j - 1)(y_i - f(\mathbf{x}_i)^{(\alpha_j)})\Big)\Bigg) \quad (5)$$

which can be seen as a sum of the MSE and the respective quantile losses, represented in one scalar. This scalar is used as the new backpropagation objective.

In Figure 3 we use a toy example to illustrate the differences between the MSE and Quantile loss: the former increases very fast in case of outliers, whereas the latter is more robust. For the individual quantiles, we observe that for very extreme values (e.g. 0.01 or 0.99) the loss skews significantly assigning high penalties to underestimation and overestimation, respectively. In our context, very athletic or sedentary people can be considered as long-tail outliers and we want our models to account for it. Intuitively, the proposed loss can be seen as a combination of multiple objective functions where the second term acts as a regularizer for the MSE. During our experiments in next sections we apply different ablations of these terms to evaluate their impact.

Feature	Seq.	Inp.	Unit
Sensor			
Acceleration	✓	✓	m/s^2
Heart Rate	✓	✗	Beats/Min. (BPM)
Timestamp	✓	✓	N/A
Metadata			
UserID	✗	✗	N/A
Height	✗	✗	Meters
Weight	✗	✗	Kilograms
Sex	✗	✗	Male–Female
Resting HR	✗	◇	BPM
VO_{2max}	✗	✗	$mL/min \cdot kg$
Derived			
Triaxial Acceleration	✓	✓	m/s^2
ENMO	✓	✓	milli-g
VM-HPF	✓	✓	milli-g
PAEE	✗	✗	$J/min \cdot kg$
Body Mass Index (BMI)	✗	✗	kg/m^2
Month, Hour	✗	◇	cos-sin transform

Table 2: Data description. *Seq.* denotes sequential measurements (timeseries), while *Inp.* the inputs to the pre-training task. (◇ feature used in some models, see Results)

4 EVALUATION

Dataset. The Fenland study is a prospective cohort study that includes 12,435 men and women who are between the ages of 35-65 [42] . After a baseline clinic visit, a subsample of 2,100 participants were asked to wear a combined heart rate and movement chest sensor and a wrist accelerometer on their non-dominant wrist. All participants provided written informed consent and the study was approved by the NRES - Cambridge East Research Ethics Committee (IRAS ID 138617).

Study protocol. The *chest ECG* measured heart rate and uniaxial acceleration in 15-second intervals while the *wrist device* recorded 60 Hz triaxial acceleration. The chest device was attached to the chest at the base of the sternum by two standard ECG electrodes. Participants were told to wear both monitors continuously 24/7 for a week and were advised that both monitors were waterproof and could be worn during showering, sleeping or exercising. During a lab visit, all participants performed a treadmill test that was used to inform their VO_2max (maximum rate of oxygen consumption and a golden measure of fitness). Resting Heart Rate (RHR) was measured with the participant in a supine position using the *chest ECG*. HR was recorded for 15 minutes and RHR was calculated as the mean heart rate measured during the last 3 minutes. These measurements were then used to calculate the Physical Activity Energy Expenditure (PAEE) [6].

Pre-processing. All participant heart rate data collected during free-living conditions underwent pre-processing for noise removal [54]. Similarly, all accelerometer data was auto-calibrated to local gravity, non-wear time was inferred and participants with less than 72 hours of wear were removed. Magnitude of acceleration was calculated through the *Euclidean Norm Minus One* (ENMO) and

the *high-passed filtered vector magnitude (VM-HPF)* (expressed in milli-g/mg per sample). Both the accelerometry and ECG signals were summarized to a common time resolution of one observation per 15 seconds and no further processing to the original signals was applied. Since the time can have a big impact on physical activity (sleeping, commuting or even the season of the year), we encoded the sensor timestamps using *cyclical temporal features* T_f [8]. Here we encoded the month of the year and the hour of the day as (x, y) coordinates on a circle:

$$T_{f_1} = sin\Big(\frac{2 * \pi * t}{max(t)}\Big) \quad (6) \qquad T_{f_2} = cos\Big(\frac{2 * \pi * t}{max(t)}\Big) \quad (7)$$

where t is the relevant temporal feature (hour or month). The intuition behind this encoding is that the model will "see" that e.g. 23:59 and 00:01 are 2 minutes apart (not 24 hours).

Training procedure. To create appropriate training batches for deep learning, we segmented the signals into fixed *non-overlapping* windows of 512 timesteps, each one comprising 15-seconds and therefore yielding a window size of approximately 2 hours. In other words, we slice the data in such a way so that the activity signals consist of a window spanning from two hours ago until the present, while the forecast heart rate is 15" *after* the last activity sample. A *sensor window*, in this case, is the result of splitting the week-long user data into smaller chunks. The resulting dataset is divided into training and test sets randomly using an 80-20% split, with the training set then being further split into training and validation sets (90-10%). We ensured that the test and train set had disjoint user groups (unseen participants are used for model evaluation). Further, we normalized the data by performing min-max scaling on all features described on Table 2 (sequence-wise for timeseries and column-wise for tabular ones) on the training set and applying it to the test set. During training, the target data (HR bpm) is not scaled and the forecast is 15" in the future after the last activity input.

Network parameters. The neural network was built through a stack of 2 CNN layers of 128 filters each, followed by 2 Bidirectional GRU stacked layers of 128 units each (resulting in 256 features due to bidirectional passes). When using extra inputs (RHR or timestamp derived features), a *ReLu* MLP of dimensionality 128 was employed for each one and its outputs were concatenated with the GRU output. We trained using the Adam [26] optimizer for 300 epochs or until the validation loss stopped improving for 5 consecutive epochs [3]. The quantiles we used were [0.01, 0.05, 0.5, 0.95, 0.99] so that they equally cover extreme and central tendencies of the heart rate distribution. The XGBoost baseline's hyperparameters were found through 5-fold cross validation and were then applied to the test set. Likewise, in the transfer learning task, we followed the same procedure for Logistic Regression.

Label and embeddings extraction. For the transfer learning task, we studied whether the learned embeddings **E** can predict user variables ranging from demographics to fitness and health. Since a slightly lower number of users (1506) had sufficient fitness data obtained from the lab test visit, we report only their results (the users remained in the same train/test splits $\tilde{\mathcal{D}}_{train}$ / $\tilde{\mathcal{D}}_{test}$ as earlier). To create binary labels we calculated the 50% percentile in each variable's distribution on the training set and assigned equally sized positive-negative classes. Therefore, even continuous outcomes such as BMI or age become binary targets for simplification purposes (the prediction is high/low BMI etc). The window-level embeddings were averaged with an element-wise mean pooling to produce user-level embeddings[4]. Then, to reduce overfitting, Principal Component Analysis (PCA) was performed on the training embeddings after standard scaling and the resulting projection was applied to the test set. We examined various cutoffs of explained variance for PCA, ranging from 90% to 99.9%. Intuitively, lower explained variance retained fewer components; in practice the number of components ranged from 10 to 160.

[3]hyper-parameter search was conducted with different layer numbers, unit sizes, learning rates and optimizers and we evaluated their impact on the validation set.

	MSE	RMSE	MAE
$Step2Heart_A$	144.61 (0.62)	12.02 (0.02)	9.23 (0.03)
$Step2Heart_{A/T}$	143.65 (0.28)	11.98 (0.01)	9.21 (0.03)
$Step2Heart_{A/R}$	91.76 (0.12)	9.57 (0.00)	6.92 (0.03)
$Step2Heart_{A/R/T}$	**91.11** (0.37)	**9.54** (0.01)	**6.88** (0.02)
Baselines			
Global mean	250.99	15.84	12.46
User mean	186.05	13.64	10.40
$XGBoost_A$	162.92 (0.20)	12.76 (0.00)	9.83 (0.00)

Table 3: Forecasting task results. Ablation test to compare the HR forecasting error using different input modalities and baselines.

	MSE	RMSE	MAE
$Step2Heart_{A/R/T}$			
$\mathcal{L}_{MSE}$	91.11 (0.37)	9.54 (0.01)	6.88 (0.02)
$\mathcal{L}_{MSE+Q}$	90.94 (1.12)	9.53 (0.05)	6.90 (0.10)
$\mathcal{L}_{0.5*MSE+Q}$	**90.27** (0.53)	**9.50** (0.02)	6.81 (0.05)
$\mathcal{L}_{Q}$	92.0 (0.16)	9.59 (0.00)	**6.75** (0.02)

Table 4: Loss function results. Ablation test to compare the best performing model with regards to different loss functions.

4.1 Baselines and metrics

For our baselines, we used naive lower bounds as well as modern ML models (similar to those used in previous works [19, 39]):

- **Convolutional Autoencoder:** A convolutional autoencoder learns to compress the input data ($\mathbf{X} \rightarrow \mathbf{X}$) with a reconstruction loss. This uni-modal baseline uses movement data only and is conceptually similar, albeit simpler, to [1, 45]. The intuition behind this choice is to assess whether *Step2Heart* learns better representations due to learning a multimodal mapping of movement to heart rate ($\mathbf{X} \rightarrow \mathbf{y}$). To make a fair comparison, it has similar number of parameters to the self-supervised models and we use the bottleneck layer to extract embeddings (128 dimensions). This baseline is used only for the transfer learning experiments.

[4]we experimented with min, max and median pooling over embeddings but yielded consistently worse results across all variables.

Outcome	AUC											
	Conv. Autoencoder				*Step2Heart*$_{A/T}$				*Step2Heart*$_{A/R/T}$			
PCA*	90%	95%	99%	99.9%	90%	95%	99%	99.9%	90%	95%	99%	99.9%
VO_2max	52.6	52.6	59.6	61.8	58.6	60	63.9	64.5	**68.3**	67.8	68	68.2
PAEE	69.6	70.0	70.2	71.8	74.7	74.7	77.5	76.8	78.2	79.2	**80.6**	79.7
Height	60.8	60.3	75.9	79.4	66	67.4	77.4	**82.1**	70.3	74	80.5	81.3
Weight	56.5	56.2	70.3	72.1	65.7	67.6	75	77.2	69.9	70.7	**77.4**	76.9
Sex	66.7	67.0	86.5	89.7	72.3	72.9	87.1	93.2	76.2	81.5	91.1	**93.4**
Age	46.2	46.3	53.9	59.5	55.0	61.7	66.2	66.9	61.1	63.8	67.3	**67.6**
BMI	51.6	51.5	60.1	61.2	62.8	63	68.2	67.6	64.7	66.1	67.8	**69.4**
Resting HR	49.1	49.4	55.8	55.4	56.7	56.6	**62.7**	61.7	N/A			

Table 5: Transfer learning results. Performance of embeddings in predicting variables related to health, fitness and demographic factors. A random baseline yields an AUC of 50. All values are ×100 for better legibility. (*percentage of explained variance by compressing the dimensionality of embeddings with PCA)

- **Gradient Boosting (XGboost)**: gradient boosting machines are among the best performing ML methods [11]. Since XGboost cannot work directly with timeseries, we extracted the following statistical features from the sensor windows: mean, std, max, min, percentiles (25%, 50%, 75%) and the slope of a linear regression fit. The final feature vector consists of 80 features.
- **Global mean:** Predicts y_i at each time step as the global HR mean of the training set. This is a naive baseline that assumes all users have the same HR anytime but provides a good lower bound for this longitudinal dataset.
- **User mean:** *Personalized* baseline obtained by predicting y_i at each time step as the mean value for all the user's **X** in the training set. This is similar to the previous baseline but considers the entire heart rate range of each user over the study week.

Given the continuous nature of the forecasting task, we employ standard evaluation metrics such as the Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE) for our evaluation. For the transfer learning task, the evaluation metric is the Area under curve (AUC).

5 RESULTS

5.1 Pre-training

We consider different ablation tests for *Step2Heart* as well as several baselines and report the average and standard deviation of 3 runs. For our ablation tests we consider the same model with different inputs: acceleration features only (A), with temporal features (A/T), with resting heart rate (A/R) and with both temporal features and resting heart rates (A/R/T).

Impact of the Resting Heart Rate. All results are summarized in Table 3. *Step2Heart* outperforms all baselines for this forecasting task and, when including temporal features and RHR (*Step2Heart* (A/R/T)), all performance metrics improve, resulting in an RMSE of 9.54. We note that the RMSE is probably the most interpretable metric since it directly translates to the error in HR beats per minute. Given the acceleration input, the addition of the RHR appears to be the most significant one, improving the RMSE by ∼ 2.5 and validating previous research that highlights RHR as a powerful bio-marker [17].

Implicit personalization. Interestingly, the baselines also reinforce the importance of personalized approaches as the user mean baseline vastly outperforms the global mean. Our models implicitly learn personalized patterns outperforming all baselines. Given the strong results of the embeddings in demographic prediction we present in the next section, we postulate that these models learn personalized features which would not be possible with other methods that –for example– require user-specific layers and might not scale in large-scale datasets [22].

Impact of the joint loss. When comparing different loss functions with the best performing model *Step2Heart*(A/R/T), we see (Table 4) that the proposed loss function better captures the long tails of HR. The lowest error, 9.5 RMSE, is achieved when weighting the MSE loss with the rest of the quantiles ($\mathcal{L}_{0.5*MSE+Q}$). Notably the pure quantile model achieves the best MAE of 6.75. We understand that a model optimized with the MSE loss would achieve better MSE score and a model including the 50% quantile would optimize the MAE score. Thus, for this experiment we evaluate the impact of the losses *across* all 3 metrics. In this case, the joint losses achieve the best results; the $\mathcal{L}_Q$ model may achieve the best MAE but predicably falls short in the other metrics. Given the overlapping standard deviations of the joint models ($\mathcal{L}_{0.5*MSE+Q}$ and $\mathcal{L}_{MSE+Q}$) we consider both to be our best models, however we pick the former as the one with the lowest average error.

5.2 Transfer learning

For this set of results, we use the best-performing model as shown above ($\mathcal{L}_{0.5*MSE+Q}$), extract embeddings and train linear classifiers for different outcomes. All results are presented in Table 5.

Effect of embeddings in generalization. Quantitatively, the embeddings achieved strong results in predicting variables like users' sex, height, PAEE and weight (0.93, 0.82, 0.80 and 0.77 AUC respectively). Also, BMI, VO_2max and age are moderately predictable (0.70 AUC). The pure acceleration model (A/T) moderately predicts

Resting HR (0.62 AUC), but this does not apply to the (A/R/T) since it already includes the RHR as input. Generally, the A/R/T model outperforms the A/T model showing that using the RHR as input is helpful, as discussed in the previous sections.

Impact of the new pre-training task. Our results validate previous studies like [19] with different and very aggregated data. As a simple baseline, we followed their idea of using the RHR as a single predictor and we could not surpass an AUC of 0.55 for BMI and age. Also, the autoencoder baseline, which learns to compress the activity data, under-performs when compared to $Step2Heart_{A/T}$, illustrating that the proposed task of mapping activity to HR captures the physiological state of the user, which translates to more generalizable embeddings. We note that both approaches operate only on activity data as inputs. This shows that the embeddings carry richer information than single biomarkers or modalities by leveraging the relationship between physical activity and heart rate responses.

Clinical relevance of results. Obtaining these outcomes in large populations can be valuable for downstream health-related inferences which would normally be costly and burdensome (for example a VO_2max test requires expensive laboratory treadmill equipment and respiration instruments). Additionally, PAEE has been strongly associated with lower risk of mortality in healthy older adults [34]. Similarly, VO_2max is prospectively associated with the incidence of type 2 diabetes [25].

Impact of the latent dimensionality size. From the representation learning perspective, we observe considerable gain in accuracy in some variables when retaining more dimensions (PCA components). More specifically, Sex and Height improve in absolute around +0.20 in AUC. However, this behavior is not evident in other variables such as PAEE and VO_2max, which seem robust to any dimensionality reduction. This means that the demographic variables leverage a bigger dimensional spectrum of latent features than the fitness variables which can be predicted with a subsample of the features. These findings could have great implications when deploying these models in mobile devices and deciding on model compression or distillation approaches [21].

Visualizing the latent space. Qualitatively, we visualized the resulting *latent* space in 2D with t-Distributed Stochastic Neighbor Embedding (t-SNE) [32] as shown in Figure 4. In this setup, we used the embeddings of the entire dataset. We found that many of the outcomes, like the depicted PAEE cluster in their own specific regions. We color code the extreme PAEE users in order to illustrate that most normal users are grouped in the center but high/low PAEEs are diametrically opposed. These visualizations can help us understand common behaviours (similar users are neighbors in the latent space), would allow for risk stratification and potentially suggest interventions to specific groups (e.g. nutrition or exercise advice to high-risk BMI–obesity onset cluster).

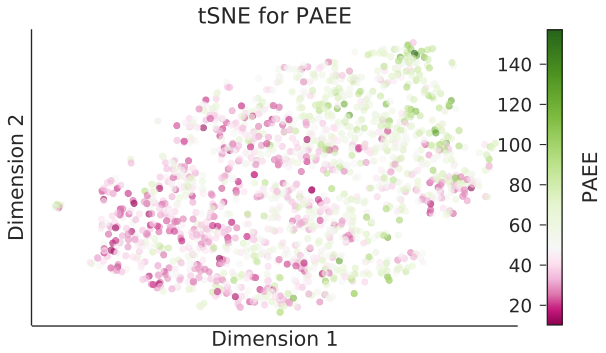

Figure 4: Model embeddings for transfer learning visualized with t-SNE. 2D representation of the test set colored with the PAEE outcome, with the colorbar showing the extreme values (the median participant has PAEE=48, white color). See Table 5 for full results.

5.3 Discussion

Our results showcase the generalizability of the proposed models to solve different tasks pertinent to physiological and behavioral data.

Viewing the upstream task in isolation (HR forecasting), we consider the error acceptable for real-world deployments, especially in cases of energy-constraint environments where the heart rate sensors could be prohibiting (an accelerometer consumes substantially less power). In our future work, we will assess the feasibility of the deployment of such model and examine its performance in different conditions (e.g. HR is generally steadier during sleep and that may affect the average error).

Further, some interesting extensions to the transfer learning experiments would be quantifying the optimal number of hours/days of data we need for each user in order to accurately predict these health-related outcomes. This could have cost saving implications as well, given that large population studies like the UK Biobank [15] or the *All of Us* [40] procure wearable devices for large cohorts over long periods of time and therefore the shortest monitoring period would be beneficial. Our current approach assumes that all temporal windows over the observation week for each user are aggregated resulting in a user-level embedding.

Zooming out, we consider the latent information captured as the most important finding of our method. Drawing parallels from the fields of natural language processing (NLP) and computer vision (CV), pioneers in representation learning [50], we posit that the behavioral and physiological signals captured by wearable sensors are appropriate and suitable for neural embeddings. In NLP and CV, researchers share pre-trained networks that can then be used to solve various downstream tasks. Inspired by the terminology used in [9], physiological signals display similar levels of *complexity* (it is not trivial to generate hand-crafted features) and *consistency* (movement is reflected as an increase in acceleration across all people) to NLP and CV. We believe that this could motivate a similar paradigm shift in the area of mobile health data especially given the privacy constraints associated with sharing such data. Instead, sharing models and embeddings would not directly expose participants' information and could accelerate research in a privacy-conscious way.

6 CONCLUSION

Here we proposed a novel *self-supervised* general-purpose neural network which can be used as a feature extractor for wearable data. These features can be used for a variety of practical downstream tasks that are *personalized* to the users' unique physiology. We evaluated this model with the largest dataset of its kind, including over 1,700 participants with combined heart and activity sensors for a week. Our model outperforms a set of strong baselines in both upstream and downstream tasks evaluated with ablation studies.

In the upstream task we found that including a single measure of RHR had significant impact, and in combination with cyclical modeling of the timestamps achieved the lowest error of ∼9 BPM in free living conditions. Nevertheless, even the model solely relying on acceleration (A/T) achieved competitive results (∼12 BPM) outperforming other ML baselines. We also introduced a joint *loss function* in order to capture the long-tails of HR observed in the real world. These joint losses outperformed single losses across all error metrics. The task-agnostic embeddings achieved strong performance at inferring physiologically meaningful variables (BMI, fitness etc), outperforming unsupervised autoencoders and common bio-markers. By inspecting the embeddings we also noticed most outcomes improve with higher latent dimensionality, while some are invariant to its size. More fine-grained prediction of the outcomes as well as comparison with contrastive approaches [12, 57] is also left for future work. Last, this method proposed hereby could potentially be applied to other domains where parallel time-series are prevalent (weather, traffic etc) in order to learn rich cross-modal representations.

7 ETHICAL AND BROADER IMPLICATIONS

The healthcare industry is undergoing an unprecedented digital transformation, producing and curating large amounts of data. Annotating all this data in order to feed to deep learning models for pattern recognition is impractical. Through self-supervised learning, we can leverage this unlabelled data to learn meaningful representations that can generalize in situations where ground truth is inadequate or simply infeasible to collect due to high costs. Such scenarios are of great importance in population health where we may be able to achieve clinical-grade health inferences with widely-adopted devices such as wearables and smartphones. Our work makes contributions in the area of transfer learning and subject-specific representations, one of utmost importance in machine learning for health.

Personalized health-representations like the ones arising from our models could raise some concerns if used maliciously for exclusionary insurance policies or unfair credit scoring, for example. However, we should clarify that our proposed model is a *tool*, and like all tools might be subject to misuse. Hence, while the risks associated to *Step2Heart* are minimal, it is paramount that future developments and use of this technology follow data governance principles that guarantee the rights of users, prevent misuse of data and promote trust in the rapidly evolving digital health ecosystem.

8 ACKNOWLEDGEMENTS

D.S was supported by the Embiricos Trust Scholarship of Jesus College Cambridge, and EPSRC through Grant DTP (EP/N509620/1). I.P was supported by GlaxoSmithKline and EPSRC through an iCase fellowship (17100053). The authors declare that there is no conflict of interest regarding the publication of this work.

REFERENCES

[1] Karan Aggarwal, Shafiq Joty, Luis Fernandez-Luque, and Jaideep Srivastava. 2019. Adversarial unsupervised representation learning for activity time-series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 834–841.
[2] Mohammad Abu Alsheikh, Ahmed Selim, Dusit Niyato, Linda Doyle, Shaowei Lin, and Hwee-Pink Tan. 2015. Deep activity recognition models with triaxial accelerometers. *arXiv preprint arXiv:1511.04664* (2015).
[3] Tim Althoff, Jennifer L Hicks, Abby C King, Scott L Delp, Jure Leskovec, et al. 2017. Large-scale physical activity data reveal worldwide activity inequality. *Nature* 547, 7663 (2017), 336.
[4] Brandon Ballinger, Johnson Hsieh, Avesh Singh, Nimit Sohoni, Jack Wang, Geoffrey H Tison, Gregory M Marcus, Jose M Sanchez, Carol Maguire, Jeffrey E Olgin, et al. 2018. DeepHeart: semi-supervised sequence learning for cardiovascular risk prediction. In *AAAI*.
[5] Davis W Blalock and John V Guttag. 2016. Extract: Strong examples from weakly-labeled sensor data. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 799–804.
[6] Søren Brage, Niels Brage, Paul W Franks, Ulf Ekelund, Man-Yu Wong, Lars Bo Andersen, Karsten Froberg, and Nicholas J Wareham. 2004. Branched equation modeling of simultaneous accelerometry and heart rate monitoring improves estimate of directly measured physical activity energy expenditure. *Journal of applied physiology* 96, 1 (2004), 343–351.
[7] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)* 46, 3 (2014), 1–33.
[8] Debaditya Chakraborty and Hazem Elzarka. 2019. Advanced machine learning techniques for building performance simulation: a comparative analysis. *Journal of Building Performance Simulation* 12, 2 (2019), 193–207.
[9] Hugh Chen, Scott Lundberg, Gabe Erion, Jerry H Kim, and Su-In Lee. 2020. Deep Transfer Learning for Physiological Signals. *arXiv preprint arXiv:2002.04770* (2020).
[10] Richard Chen, Filip Jankovic, Nikki Marinsek, Luca Foschini, Lampros Kourtis, Alessio Signorini, Melissa Pugh, Jie Shen, Roy Yaari, Vera Maljkovic, et al. 2019. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In *KDD*.
[11] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *KDD*.
[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
[13] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP* (2014).
[14] Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. 2018. Distributional reinforcement learning with quantile regression. In *AAAI*.
[15] Aiden Doherty, Dan Jackson, et al. 2017. Large scale population assessment of physical activity using wrist worn accelerometers: the UK biobank study. *PloS one* 12, 2 (2017), e0169649.
[16] MYRVIN H Ellestad and MKE Wan. 1975. Predictive implications of stress testing. Follow-up of 2700 subjects after maximum treadmill stress testing. *Circulation* 51, 2 (1975), 363–369.
[17] Kim Fox, Jeffrey S Borer, A John Camm, Nicolas Danchin, Roberto Ferrari, Jose L Lopez Sendon, Philippe Gabriel Steg, Jean-Claude Tardif, Luigi Tavazzi, Michal Tendera, et al. 2007. Resting heart rate in cardiovascular disease. *Journal of the American College of Cardiology* 50, 9 (2007), 823–830.
[18] David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122* (2018).
[19] Haraldur T Hallgrímsson, Filip Jankovic, Tim Althoff, and Luca Foschini. 2018. Learning Individualized Cardiovascular Responses from Large-scale Wearable Sensors Data. *NIPS ML4H workshop* (2018).
[20] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. *IJCAI* (2016).
[21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).
[22] Natasha Jaques, Sara Taylor, Akane Sano, Rosalind Picard, et al. 2017. Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation. In *IJCAI 2017 Workshop on artificial intelligence in affective computing*. 17–33.
[23] Simon Jenni and Paolo Favaro. 2018. Self-supervised feature learning by learning to spot artifacts. In *CVPR*.

[24] Andrew M Jones and Helen Carter. 2000. The effect of endurance training on parameters of aerobic fitness. *Sports medicine* 29, 6 (2000), 373–386.

[25] Peter T Katzmarzyk, Timothy S Church, Ian Janssen, Robert Ross, and Steven N Blair. 2005. Metabolic syndrome, obesity, and mortality: impact of cardiorespiratory fitness. *Diabetes care* 28, 2 (2005), 391–397.

[26] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[27] Adit Krishnan, Ashish Sharma, and Hari Sundaram. 2018. Insights from the long-tail: Learning latent representations of online user behavior in the presence of skew and sparsity. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 297–306.

[28] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *ICLR* (2020).

[29] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*.

[30] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2017. Unsupervised representation learning by sorting sequences. In *ICCV*.

[31] Haojie Ma, Wenzhong Li, Xiao Zhang, Songcheng Gao, and Sanglu Lu. 2019. AttnSense: multi-level attention mechanism for multimodal human activity recognition. In *IJCAI*.

[32] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[33] Kyle Mandsager, Serge Harb, Paul Cremer, Dermot Phelan, Steven E Nissen, and Wael Jaber. 2018. Association of cardiorespiratory fitness with long-term mortality among adults undergoing exercise treadmill testing. *JAMA network open* 1, 6 (2018), e183605–e183605.

[34] Todd M Manini, James E Everhart, Kushang V Patel, Dale A Schoeller, Lisa H Colbert, Marjolein Visser, Frances Tylavsky, Douglas C Bauer, Bret H Goodpaster, and Tamara B Harris. 2006. Daily activity energy expenditure and mortality among older adults. *Jama* 296, 2 (2006), 171–179.

[35] Junhua Mao, Jiajing Xu, Kevin Jing, and Alan L Yuille. 2016. Training and evaluating multimodal word embeddings with large-scale web annotated images. In *Advances in neural information processing systems*. 442–450.

[36] Ryan McConville, Gareth Archer, Ian Craddock, Herman ter Horst, Robert Piechocki, James Pope, and Raul Santos-Rodriguez. 2018. Online heart rate prediction using acceleration from a wrist worn wearable. *arXiv preprint arXiv:1807.04667* (2018).

[37] Mehdi Menai, Vincent T Van Hees, Alexis Elbaz, Mika Kivimaki, Archana Singh-Manoux, and Séverine Sabia. 2017. Accelerometer assessed moderate-to-vigorous physical activity and successful ageing: results from the Whitehall II study. *Scientific reports* 7 (2017), 45772.

[38] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *ICML*.

[39] Jianmo Ni, Larry Muhlstein, and Julian McAuley. 2019. Modeling Heart Rate and Activity Data for Personalized Fitness Recommendation. In *WWW*.

[40] All of Us Research Program Investigators. 2019. The "All of Us" Research Program. *New England Journal of Medicine* 381, 7 (2019), 668–676.

[41] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. 2016. Ambient sound provides supervision for visual learning. In *ECCV*.

[42] Laura O'Connor, Soren Brage, Simon J Griffin, Nicholas J Wareham, and Nita G Forouhi. 2015. The cross-sectional association between snacking behaviour and measures of adiposity: the Fenland Study, UK. *British journal of nutrition* 114, 8 (2015), 1286–1293.

[43] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. 2018. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–27.

[44] Filipe Rodrigues and Francisco C Pereira. 2018. Beyond expectation: Deep joint mean and quantile regression for spatio-temporal problems. *arXiv preprint arXiv:1808.08798* (2018).

[45] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-task Self-Supervised Learning for Human Activity Detection. *IMWUT* 3, 2 (2019), 61.

[46] Benjamin Sanchez-Lengeling, Jennifer N Wei, Brian K Lee, Richard C Gerkin, Alán Aspuru-Guzik, and Alexander B Wiltschko. 2019. Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules. *arXiv preprint arXiv:1910.10685* (2019).

[47] Pritam Sarkar and Ali Etemad. 2019. Self-supervised Learning for ECG-based Emotion Recognition. *arXiv preprint arXiv:1910.07497* (2019).

[48] Kai P Savonen, Timo A Lakka, Jari A Laukkanen, Pirjo M Halonen, Tuomas H Rauramaa, Jukka T Salonen, and Rainer Rauramaa. 2006. Heart rate response during exercise test and cardiovascular mortality in middle-aged men. *European heart journal* 27, 5 (2006), 582–588.

[49] Patrick Schwab and Walter Karlen. 2019. PhoneMD: Learning to diagnose Parkinson's disease from smartphone data. In *AAAI*.

[50] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging* 35, 5 (2016), 1285–1298.

[51] Dimitris Spathis, Ignacio Perez-Pozuelo, Soren Brage, Nicholas J Wareham, and Cecilia Mascolo. 2020. Learning Generalizable Physiological Representations from Large-scale Wearable Data. *arXiv preprint arXiv:2011.04601* (2020).

[52] Dimitris Spathis, Sandra Servia-Rodriguez, Katayoun Farrahi, Cecilia Mascolo, and Jason Rentfrow. 2019. Passive mobile sensing and psychological traits for large scale mood prediction. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 272–281.

[53] Dimitris Spathis, Sandra Servia-Rodriguez, Katayoun Farrahi, Cecilia Mascolo, and Jason Rentfrow. 2019. Sequence multi-task learning to forecast mental wellbeing from sparse self-reported data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2886–2894.

[54] Oliver Stegle, Sebastian V Fallert, David JC MacKay, and Søren Brage. 2008. Gaussian process robust regression for noisy heart rate data. *IEEE Transactions on Biomedical Engineering* 55, 9 (2008), 2143–2151.

[55] Hirofumi Tanaka, Kevin D Monahan, and Douglas R Seals. 2001. Age-predicted maximal heart rate revisited. *Journal of the american college of cardiology* 37, 1 (2001), 153–156.

[56] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, Soren Brage, Nick Wareham, and Cecilia Mascolo. 2021. SelfHAR: Improving Human Activity Recognition through Self-training with Unlabeled Data. *arXiv preprint arXiv:2102.06073* (2021).

[57] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, and Cecilia Mascolo. 2020. Exploring Contrastive Learning in Human Activity Recognition for Healthcare. *arXiv preprint arXiv:2011.11542* (2020).

[58] Xian Wu, Chao Huang, Pablo Roblesgranda, and Nitesh Chawla. 2020. Representation Learning on Variable Length and Incomplete Wearable-Sensory Time Series. *arXiv preprint arXiv:2002.03595* (2020).

[59] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. 2015. Deep convolutional neural networks on multichannel time series for human activity recognition. In *IJCAI*.

[60] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. 2017. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *WWW*.