Machine learning to model health with multimodal mobile sensor data

Dimitrios Spathis

Jesus College
University of Cambridge

July 2021

This dissertation is submitted for the degree of *Doctor of Philosophy* at the Department of Computer Science and Technology

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Dimitrios Spathis

Abstract

Machine learning to model health with multimodal mobile sensor data

Dimitrios Spathis

The widespread adoption of smartphones and wearables has led to the accumulation of rich datasets, which could aid the understanding of behavior and health in unprecedented detail. At the same time, machine learning and specifically deep learning have reached impressive performance in a variety of prediction tasks, but their use on time-series data appears challenging. Existing models struggle to learn from this unique type of data due to noise, sparsity, long-tailed distributions of behaviors, lack of labels, and multimodality. This dissertation addresses these challenges by developing new models that leverage multi-task learning for accurate forecasting, multimodal fusion for improved population subtyping, and self-supervision for learning generalized representations. We apply our proposed methods to challenging real-world tasks of predicting mental health and cardio-respiratory fitness through sensor data.

First, we study the relationship of passive data as collected from smartphones (movement and background audio) to momentary mood levels. Our new training pipeline, which combines different sensor data into a low-dimensional embedding and clusters longitudinal user trajectories as outcome, outperforms traditional approaches based solely on psychology questionnaires. Second, motivated by mood instability as a predictor of poor mental health, we propose encoder-decoder models for time-series forecasting which exploit the bi-modality of mood with multi-task learning.

Next, motivated by the success of general-purpose models in vision and language tasks, we propose a self-supervised neural network ready-to-use as a feature extractor for wearable data. To this end, we set the heart rate responses as the supervisory signal for activity data, leveraging their underlying physiological relationship and show that the resulting task-agnostic embeddings can generalize in predicting structurally different downstream outcomes through transfer learning (e.g. BMI, age, energy expenditure), outperforming unsupervised autoencoders and biomarkers. Finally,

acknowledging fitness as a strong predictor of overall health, which, however, can only be measured with expensive instruments (e.g., a VO_2max test), we develop models that enable accurate prediction of fine-grained fitness levels with wearables in the present, and more importantly, its direction and magnitude almost a decade later.

All proposed methods are evaluated on large longitudinal datasets with tens of thousands of participants in the wild. The models developed and the insights drawn in this dissertation provide evidence for a better understanding of high-dimensional behavioral and physiological data with implications for large-scale health and lifestyle monitoring.

Acknowledgements

Pursuing a PhD can be a lonely endeavor, but I am grateful to have had the support of brilliant mentors, friends, and family. This is for you.

First, to Cecilia Mascolo, my advisor who believed in me, and invested time, energy, and resources into helping me grow professionally and personally. Through giving me space to explore my own research directions, supernatural email responsiveness, and an incredible eye for important and impactful problems, her relentless support has turned my half-baked ideas into competent research. Through the Mobile Systems Group, Cecilia has created an amazing research team and community that I will cherish forever.

My co-advisor, Jason Rentfrow, provided this thesis with a different perspective beyond Computer Science, always had time to discuss ideas, and offered incredibly useful pointers. He also introduced me to David Greenberg, and I consider both as two of my key collaborators for the years to come.

I owe a lot to two colleagues we can even say we've been in the trenches together! Sandra Servia taught me how to be systematic and meticulous as well as that the art of writing is in the rewriting. She was crucial for the course of my PhD and our work gave me the early confidence I needed to complete this journey. Ignacio Perez-Pozuelo introduced me to the wondrous world of medical research and is due to him that the second half of my PhD has been the most productive time of my life so far. I will never forget our late evening pizzas at the Alan Turing Institute criticizing papers and debugging models! Apart from turning into a great friendship, Ignacio was the key person in our recent collaborations with MIT, Oxford, and the MRC Epidemiology Unit at Cambridge.

To my fellow co-authors Kate Farrahi, Soren Brage, Nicholas Wareham, and Tomas Gonzales, thank you for the guidance and the stimulating discussions.

I was extremely fortunate to have great academic mentors even before the PhD. Spyros Sioutas, Katia Kermanidis, and Panagiotis Vlamos put a lot of trust in me despite my young age during my undergrad years and shaped my career. Likewise, Anastasios Tefas and Athena Vakali, let me go deep into research topics I enjoyed during my master's and introduced me to international research. Further, I wouldn't

be writing these lines if it wasn't for Ilias Leontiadis. While showing me that research work can be both productive and fun, he inadvertently convinced me to pursue a PhD and eventually became my "academic brother".

I want to express my gratitude to my examination committee, Professors Mateja Jamnik and David Clifton, for their patience especially after reading the next several hundred pages, and their thoughtful suggestions that made this work even better.

During my internship at Ocado, I had the wonderful opportunity to be mentored by some of the kindest and smartest people. Laurent Candillier, my host, introduced me to the challenges of industrial-scale recommender systems, while the rest of the colleagues (Anton Malinovskiy, Mihai Ratiu, and Jose Jimenez) gave me the freedom to explore projects that were both important and foundational to my thesis topic.

To the undergraduate students that worked with me closely, in particular, Benjamin Searle, Chuen Low, and Kevalee Shah, thank for you giving me the opportunity to explore new directions, and I am excited to see what you all do in the future.

Our Group feels like a big extended family to me due to all these amazing people such as Andrea, Andreas, Alessandro, Api, David, Dionysis, Erika, Ian, Krittika, Lorena, Petko, Tassos, Xiao, and Young. I will never forget our everyday fun, and countless potlucks, trips, and outings. Science is not done in a vacuum, and I am really thankful to my friends in Cambridge and abroad who were there to raise my spirits. In particular, Stefanos, Christos, Thanasis, Minos, Sebastian, Jose, Giorgia, Ross, and Frida, were always there for me. Vasilis and Andreas might have been very far but we always managed to pick it up from where we last left it off. A special mention must be made to my bandmates Charis, Chris, Tanuj, and Phoebe who showed me that Cambridge still got some rock 'n' roll.

The Department of Computer Science and Technology at the University of Cambridge is a special place to do research, including the friendly administrative staff and my year's PhD cohort. Special shout-out to Lise Gough and Marketa Green. Also, I am grateful to receive financial support from the Department, the EPSRC through Grant DTP (EP/N509620/1), and the Embiricos Trust Scholarship of Jesus College Cambridge.

Besides mentors, colleagues, and friends, I need to give a final special thank you to my family and siblings. My parents, Antonia and Charalambos, have always prioritized my education and supported every choice of mine towards discovering my passions even so far from home. I miss you.

Finally, Laia, the love of my life, was always there to remind me what is important in life. Thank you for sticking with me through every paper deadline and stressful presentation. You believed in me even when we were thousands of miles apart. This PhD feels as much yours as it is mine. *T'estimo*.

Contents

Chapter 1

Introduction

The purpose of computation is insight, not numbers
–Richard Hamming

1.1 Motivation

Computers and data have changed the way we organise information, the way we communicate, and the way we think about science. The curation of large datasets has revolutionized many areas, enabling advances on a scale unthinkable decades ago (Aad et al., 2012). But data seen in isolation has no meaning; our efforts should be targeted to extract actionable insights and knowledge that influence decisions and ultimately improve lives.

Health data is the best candidate to directly transform lives. Advances in the ways in which we process this data can transform our society. Although the overwhelming majority of medical research studies clinical data (labs, imaging, vitals etc), the average person visits a doctor only around 5 times a year (Kim et al., 2014). On the other hand, recent advances in wearable sensing and mobile computing, along with their wide adoption, have created new pathways for the collection of health and well-being data outside of laboratory and hospital settings, in a longitudinal fashion. Apart from *"filling the gaps"* of traditional clinical data, these devices open up new research and commercial directions for large-scale lifestyle monitoring. For example, millions of people worldwide use such devices to track their physical activity and sleep, with increasingly more sophisticated predictive capabilities (Althoff et al., 2017).

At the same time, seemingly disparate forces like mature open-source scientific software libraries, easier data crowdsourcing and labeling, and the repurposing of specialized hardware (graphics cards), have enabled dramatic improvements in predictive modeling. Many machine learning (ML) tasks have achieved impressive performance,

ranging from object recognition in images (He et al., 2016), to winning the best players in the games of Go, Atari, and Chess (Schrittwieser et al., 2020), or outperforming experts in breast cancer screening (McKinney et al., 2020). The common denominator in all these cases has been the curation of high-quality large datasets that allow models to exploit latent patterns and subsequently generalize in the real world (Hyland et al., 2020). However, especially in health where erroneous predictions can have grave consequences, the roll-out and adoption of such systems has been met with resistance (Davenport and Kalakota, 2019). Instead, fields with low false-positive costs and high digitization rates such as online services, social networks, or streaming services, have not only embraced machine learning, but also actively drive the research community in further developing the fields of computer vision and natural language processing.

Similar to how social networks learn our online behaviors, wearable and mobile devices monitor our activities in the real world. By tracking our sleep, steps, eating and working habits, they create a holistic understanding of the most important components of our everyday health (World Health Organization, 2002), until now only possible through surveys. Although we recognize the value of such datasets, advances in machine learning for health and mobile sensing[1] have not kept up with other areas. For example, over the last decade devices such as Fitbit or the iPhone have been collecting multimodal sensor data at an unprecedented temporal resolution. However, effectively leveraging these datasets presents many challenges, leading to this data being frequently overlooked for scientific and medical research. Further, obtaining quality annotations and ground truth might be costly or even impossible at this granularity. New computational methods are needed to address these challenges and this thesis attempts to bridge some of these gaps.

1.2 Limitations of traditional mental and physical health monitoring

Despite the importance of detecting and understanding fluctuations of mental and physical health, physicians and researchers are hindered by a key limitation: the lack of reliable and meaningful data. Most established research and clinical practice is based on pen-and-paper self-reports and surveys which, whilst valuable in the absence of alternatives, are subject to bias and often provide incomplete information (Brenner and DeLamater, 2014).

[1] We define *mobile sensing* as data from connected sensors which is used to characterize behaviors related to health. Other terms used in literature are *personal sensing*, *digital phenotyping*, and *context sensing*, with different albeit overlapping connotations. We point the interested reader to this discussion (Mohr et al., 2020).

Individuals may inaccurately recall their behavior, report an idealised version of their habits or some combination thereof. Previous studies have found that self-reported physical activity suffers from reporting bias, which stems from social desirability bias (reporting behavior seen as socially desirable), as well as the cognitive complexity of reporting the intensity, duration, and frequency of physical activity behaviors with precision (Sallis and Saelens, 2000). In addition, the understanding of a behavior that is self-reported is limited to the specific set of questions given to study participants. These may not be enough to reflect a complete view of complex behaviors. Inaccuracies resulting from reporting errors may be randomly distributed across the population being studied. The errors may also be systematic, with participants in different population groups systematically under or over-reporting their activity levels. This could lead to the identification of erroneous associations.

Similar to physical health, bias can impact mental health studies in more subtle ways. Patients who are asked to report their mood levels or test for depression might be triggered by the content of the questions in a self-reinforcing loop that can possibly do more harm than good (Labott et al., 2013). In order to diminish concerns regarding bias in studies using self-report measures of physical and mental health, questionnaires should be validated against a gold-standard measure or objective measurements when this is possible. Data from mobile devices can combine the best of both worlds: self-reports are always time-stamped and contextual due to push notifications, while passive sensors can unobtrusively and objectively monitor behaviors.

1.3 Challenges in multi-sensory machine learning

The typical workflow of a scientist in most fields involves devising comprehensive variables[2] that explain the variance of a dataset. Until recently, this process was characterized by elaborate *feature engineering* in order to construct informative features that would discriminate between some classes (in the case of classification). Now, deep neural networks promise to automate this task by learning latent features as the side-effect of the optimization process and besides achieving state-of-art results (LeCun et al., 2015). This is even more crucial for mobile sensing data (Fig. 1.1 presents a schematic of a typical machine learning workflow for mobile sensor data).

Data coming from common sensors such as accelerometers, electrocardiograms (ECGs), gyroscopes, and microphones is commonly represented as high-dimensional time-series (Lane et al., 2010). Unlike other data types, these sensor measurements are

[2]In this thesis, we use the terms variable, feature, and covariate interchangeably, when referring to the input data of a statistical model. The ML community prefers the more liberal term *feature* which tends to describe raw data (in 3D or higher dimensions), whereas in statistics the independent variables are often the outcome of some initial processing.

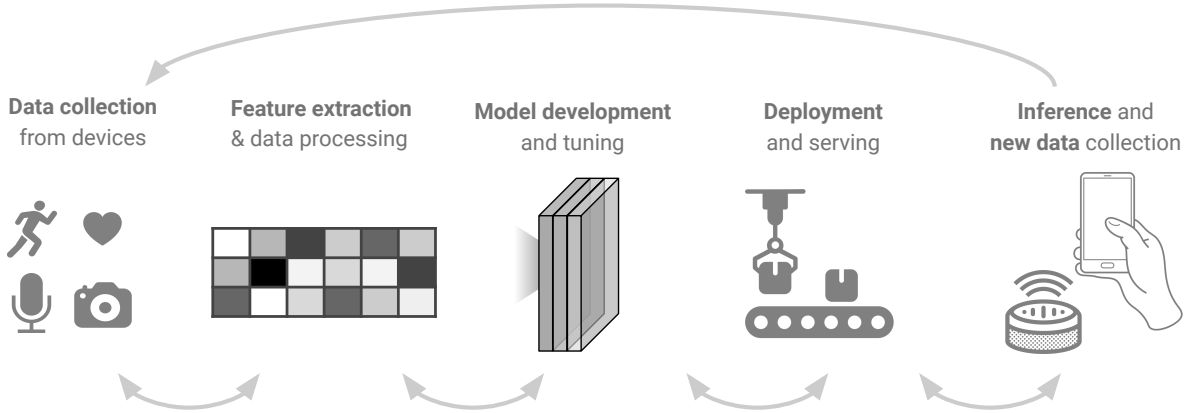


Figure 1.1: **Lifecycle of machine learning workflows**. The iterative steps followed when processing data from personal devices. This thesis makes contributions to methods applied across the entire lifecycle, with a particular focus on novel feature extraction and *representation* methods, as well as *generalizing to new data* collected in a longitudinal fashion.

noisy and although in small-scale studies manually engineered features have proved to be effective (Wang et al., 2014), it is not straightforward to select robust features for different noise levels of individual user behaviors. Noise in mobile measurements is hard to model because it is correlated over time (Park, 2004) and presents a non-linear structure (Ang et al., 2007). Apart from noise, challenges in modeling sensor data extend to varying sparsity levels (Abedin et al., 2019; Ghassemi et al., 2015), the inability to obtain quality annotations and labels (Bulling et al., 2014), and heterogeneous data types (Radu et al., 2018), unlike those used in established benchmark tasks.

An illustrative scenario depicting these challenges may be an individual taking off their smartwatch when having a shower. The light sensor of the watch might reflect off a distant surface and hence record a faulty heart rate (noise); the non-wear period produces irregularly sampled data which has to be imputed with the expected sensor values (sparsity); it is unlikely that the user would annotate this or other events at the minute level (label scarcity). Furthermore, when finally worn, motion and heart sensors behave differently in response to external stimuli such as stress (multimodality) (Bent et al., 2020).

Scale is also different. Large longitudinal studies like the *UK Biobank* (Doherty et al., 2017), the Apple Study (Perez et al., 2019), the *Fenland Study* (O'Connor et al., 2015), *Utsureko* (Suhara et al., 2017), and *EmotionSense* (Servia-Rodríguez et al., 2017) have been monitoring the physical and mental health of tens of thousands of participants with wearable sensors. For example, elevated resting heart rate from over 200, 000 Fitbit users was used to predict influenza-like illness in the US (Radin et al., 2020). However, statistical methods like generalized linear mixed models which operate on

longitudinal data with repeated measurements (e.g. a single user appears on multiple days), cannot scale to that number of subjects (Zhang et al., 2016). Also, given that previous studies in mobile health have been conducted through controlled experiments and a limited number of participants (Sano, 2016; Jaques et al., 2017; LiKamWa et al., 2013), it is not clear whether previous findings and methodologies could be transferred to these large datasets. Furthermore, the absence of rigid control over participation and the limited mechanisms to promote engagement, make the data collected more difficult to interpret than in controlled setups.

Arguably, the underlying challenge here is about *representation*. Machine learning attempts to find meaningful representations which will transform data to linearly separable spaces and distinguish between semantic classes. It has come a long way from the low-dimensional embeddings of convolutional networks that captured the structure of image datasets (LeCun et al., 2015) to the recent *self-supervised* networks which learn better features by predicting distorted samples of the input data (Devlin et al., 2019). But understanding how multi-sensory measurements relate to each other across time means constructing a representation of the *health state* of the individual. This thesis argues that some of the fundamental building blocks of future machine learning for health will be *multi-tasking*, *transfer learning*, and *forecasting*. We now know that models that perform multiple tasks are not only more useful, but they also make each individual task more robust (Kaiser et al., 2017). Also, models that are able to reason about the future can transfer better across different environments (Chen et al., 2021). Therefore, we need models to account for these challenges through improved data representation that leverages cross-sensor relationships and relying as little as possible on manual annotations.

1.4 Thesis and substantiation

We have reviewed some potential benefits from the improvement of machine learning for mobile health, what limitations arise when traditional methods are employed, and what challenges are involved when processing noisy sensor data. Formally, the overarching objective can be stated as: *to improve machine learning methods for observational, retrospective, and longitudinal data as generated by consumer mobile and wearable devices, both of dynamic and static nature, of multiple tasks, and of limited supervision, for the ultimate aim of improving health and well-being with a focus on mental and physical health*. We substantiate this statement by first evaluating the potential of existing approaches on large-scale physical and mental health datasets and then proposing new models which outperform current methods or offer new insights. Our methods leverage and expand on the paradigms of multimodal fusion, multi-task learning,

time-series forecasting, transfer learning, and self-supervised learning. In particular, this dissertation addresses the following four research questions:

- **Research Question 1**. How can we use machine learning to combine passive sensor time-series with traditional user-level metadata to distinguish between clustered user trajectories?
- **Research Question 2**. How effective is multi-task learning and encoder-decoder models for multi-step time-series forecasting?
- **Research Question 3**. How can we train general-purpose neural networks with self-supervision to leverage large amounts of unlabelled time-series data?
- **Research Question 4**. How can we use deep learning on free-living wearable sensor data for long-term cardio-respiratory fitness prediction?

To address these questions we develop models which can fuse time-series and tabular data, as well as sequence models which enable accurate forecasting of mental health. Further, we design novel self-supervised tasks that leverage large unlabeled time-series data and propose new models to predict lab-measured fitness levels with free-living sensor data.

1.5 Contributions and chapter outline

In terms of methods, we customize deep learning models to unlabeled time-series. In the application domain, we adapt machine-learning methods to challenging tasks from the fields of physical and mental health. We shall start with an introduction to the background of deep learning for sequence data in Chapter 2, before presenting the four main contributions that extend to the later chapters in the rest of the thesis as follows:

Contribution 1: Multimodal machine learning for large scale mood prediction

In Chapter 3, we show that psychological traits combined with passively collected sensing data (activity from the accelerometer and noise levels from the microphone), can detect individuals whose general mood deviates from the common relaxed characteristic. We validate our approach with data from the *EmotionSense Study*, a large mobile application dataset collected in the wild with 17,251 participants, finding that the combination of these modalities achieves the best classification performance, and

that passive sensing yields a +5% boost in accuracy. The main motivation behind this study was that experience sampling has been proposed as a mechanism to monitor mental health, but it requires users' attention and this therefore leads to considerable retention issues. We study whether passive sensing and one-off surveys can be used to identify relaxed and non-relaxed users and –by extension– unobtrusively monitor mental health.

The proposed methodology involves two steps. *First*, clustering historical mood trajectories (after feature extraction) using a standard algorithm such as k-means in order to find groups of users with similar trajectories. *Second*, classifying users into the found clusters. Our pipeline employs feature selection, dimensionality reduction and classification algorithms such as Gradient Boosting Trees and Deep Neural Networks.

The experimental results show that by adding passive sensing to personality and demographics surveys we can predict the mood group of individual users with higher precision. Our models achieve a 75% AUC when using a combination of weekly sensors (accelerometer and microphone) and one-off questionnaire data as inputs. We discuss feature extraction techniques and appropriate classifiers for this kind of multimodal data, as well as overfitting shortcomings deep neural networks when handling static and dynamic features. These findings might have significant implications for mobile health applications that can benefit from the correct modelling of passive sensing alongside extra user metadata.

Contribution 2: Multi-task and sequence learning for mood time-series forecasting

In Chapter 4, we propose an end-to-end encoder-decoder model to forecast sequences of future mood from previous self-reported mood. Our results show that multi-tasking learns both dimensions of mood simultaneously, which is more accurate than individual models or baselines. Also, plotting the neural activations helps us understand the latent trajectories of mood, as well as post-hoc error analysis identifies significant differences in the model's performance regarding the users' personality, mood variability or day of the week. The main motivation behind this study was that psychologists use mostly pen-and-paper surveys to track mental health, which, unlike mobile apps, are prone to recall bias. On a more technical side, we show that current machine learning models for mental health do not provide long-term predictions and cannot learn complex patterns from time-series.

The proposed methodology relies on an end-to-end Long Short-Term Memory (LSTM) Encoder-Decoder model. The sequence passes through an LSTM, gets transformed into a single vector, and is decoded through another LSTM that predicts future

sequences. Moreover, model interpretability is always important when dealing with health data, therefore we analyze the role of the layers of the trained models. As we move into deeper layers, we see that the network lays out the continuum of positive and negative mood, even though it has been trained to solely forecast the mood. Also, by inspecting individual neurons of the Decoder we observe that some neurons fire almost always with the same slope, while others are more conservative with almost flat lines. This helps us identify different subtypes of mood evolution.

Here, we again use data from the *EmotionSense Study*, however, this time we only use the sequences of self-reported mood. Our results show that 3 weeks is the best window of mood reporting, verifying previous research on depression prediction. Also, our models outperform machine learning regressors and simple baselines while multi-task learning seems to help the prediction of the alertness (one of the two mood dimensions). We believe this work provides psychologists and developers of future mobile mental health applications with a ready-to-use and effective tool for early diagnosis of mental health issues at scale.

Contribution 3: Self-supervised transfer learning of physiological representations from free-living wearable data

In Chapter 5, we develop a novel self-supervised general-purpose neural network which maps activity data to heart rate responses and can be used as a feature extractor for wearable data. Its features can be used for a variety of practical downstream tasks that are personalized to the users' unique physiology as well as this model outperforms a set of strong baselines in both upstream and downstream tasks evaluated with ablation studies.

For pre-training, we introduce a joint loss function that acts as a regularizer to traditional Mean Squared Error 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. There, we show that including a single measure of Resting Heart Rate had significant impact, and in combination with cyclical modeling of the timestamps achieved the lowest error of $\sim$9 BPM in free living conditions.

Downstream, 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 biomarkers. 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. By inspecting the embeddings we also notice that most outcomes improve with higher latent dimensionality, while some

are invariant to its size.

We evaluate this model with the *Fenland Study*, the largest multimodal wearable ECG and wrist accelerometry dataset, including over 1,700 participants tracked for a week, together 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. Overall, we propose a multimodal self-supervised method for behavioral and physiological data with implications for large-scale health and lifestyle monitoring

Contribution 4: Adaptable cardio-respiratory fitness prediction from free-living wearable devices using deep learning

In Chapter 6, we develop deep learning models utilising wearable data and common biomarkers to predict the gold standard of fitness (VO_2max) and achieve strong performance compared to other traditional approaches.

Cardio-respiratory fitness is a well-established predictor of metabolic disease and mortality. Fitness is directly measured as maximal oxygen consumption (VO_2max), or indirectly assessed using heart rate response to a standard exercise test. However, such exercise testing is costly and burdensome, limiting its utility in healthcare and large-scale population studies. Fitness can also be approximated using RHR and self-reported exercise habits but accuracy is low compared to estimates based on dynamic data. Modern wearables capture non-standardised dynamic data which could improve fitness prediction.

Here, we use a bigger cohort of the *Fenland Study* and analyze movement and heart rate signals from wearable sensors in free-living conditions from a population study comprising 11,059 participants who also underwent a standard exercise test. We develop a deep neural network model that leverages sensor information to predict VO_2max, yielding a Pearson correlation of r = 0.82 [CI 0.80-0.83], when compared to the ground truth in a holdout sample. This model outperforms conventional non-exercise fitness models and traditional biomarkers using measurements of normal daily living without the need to undertake a specific exercise test. Additionally, we show the adaptability and applicability of this approach for detecting fitness change over time in a longitudinal subsample (n = *2,675*) who repeated measurements after 7 years. We evaluate the inference capabilities of the model in the difference (delta) between the present and future fitness. For this last task, the model produced outcomes that translated to a 0.57 correlation between the delta of predicted and delta of true VO_2max. Last, the latent representations that arise from this model pave the way for fitness - aware monitoring and interventions at scale.

The last chapter of this thesis (Chapter 7) reflects on the new insights and results presented on the previous chapters and outlines limitations along with potential future research directions.

1.6 List of publications

During my Ph.D. studies, I was fortunate to establish several fruitful collaborations with computer scientists, engineers, psychologists, epidemiologists, and other domain-experts, which have yielded publications both in machine learning methods and their applications to mobile health. In particular, Chapter 3 draws from a study published in *PervasiveHealth 2019* (Spathis et al., 2019), Chapter 4 is based on a paper at *KDD 2019* (Spathis et al., 2019), Chapter 5 builds on recently published papers at *CHIL 2021* (Spathis et al., 2021) and *NeurIPS 2020 ML4MH* (Spathis et al., 2020), and last, Chapter 6 is based on to-be-submitted work. Beyond that, I co-authored some other works in the wider area of machine learning and data science, which, while not directly related to this dissertation, have nonetheless influenced my ideas.

Works related to this dissertation

- **Spathis, D**., Pozuelo, I., Brage, S., Wareham, N., & Mascolo, C. (2021). Self-supervised transfer learning of physiological representations from large scale free-living wearable data. *ACM Conference on Health, Inference, and Learning (CHIL)*, Virtual event, USA. https://doi.org/f6tt

- **Spathis, D**., Pozuelo, I., Brage, S., Wareham, N., & Mascolo, C. (2020). Learning Generalizable Physiological Representations from Large-scale Wearable Data. *Advances in Neural Information Processing Systems (NeurIPS-W), Machine Learning for Mobile Health workshop*, Virtual event, Canada. https://arxiv.org/abs/2011.04601

- **Spathis D**., Servia-Rodriguez, S., Farrahi, K., Mascolo, C., & Rentfrow, J. (2019). Sequence Multi-task Learning to Forecast Mental Wellbeing from Sparse Self-reported Data. *ACM International Conference on Knowledge Discovery and Data Mining (KDD),* Anchorage, USA. (Oral presentation, top 6% of submissions) http://doi.org/gf7nbh

- **Spathis, D**., Servia-Rodriguez, S., Farrahi, K., Mascolo, C., & Rentfrow, J. (2019). Passive mobile sensing and psychological traits for large scale mood prediction. *International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, Trento, Italy. http://doi.org/c7hk

- **Spathis*, D.**, Pozuelo*, I., Gonzales, T., Brage, S., Wareham, N., & Mascolo, C. Adaptable Cardiorespiratory Fitness Prediction from Free-living Wearable Devices Using Deep Learning (under review). *equal contribution

Other works and preprints

- **Spathis, D.**, Perez-Pozuelo, I., Marques-Fernandez, L., & Mascolo, C. (2022). Breaking away from labels: the promise of self-supervised machine learning in intelligent health. *Cell Patterns*, in press.

- Dang, T., Han*, J., Xia*, T., **Spathis, D.**, Bondareva, E., Brown, C., Chauhan, J., Grammenos, A., Hasthanasombat, A., Cicuta, P., Floto, A., & Mascolo, C. (2022). COVID-19 Disease Progression Prediction via Audio Signals: A Longitudinal Study. *arxiv preprint*. *equal contribution https://arxiv.org/abs/2201.01232

- Xia*, T., **Spathis*, D.**, Brown, C., Grammenos, A., Han, J., Hasthanasombat, Bondareva, E., Chauhan, J., Dang, T., Floto, A., Cicuta, P., & Mascolo, C. (2021). COVID-19 Sounds: A Large-Scale Audio Dataset for Digital Respiratory Screening. *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks track*. *equal contribution https://openreview.net/forum?id=9KArJb4r5ZQ

- Shah, K., **Spathis, D.**, Tang, I., & Mascolo, C. (2021). Evaluating Contrastive Learning on Wearable Timeseries for Downstream Clinical Outcomes. *Machine Learning for Health (ML4H)*. https://arxiv.org/abs/2111.07089

- Laskaridis, S., **Spathis, D.**, & Almeida, M. (2021). Federated mobile sensing for activity recognition. *International Conference on Mobile Computing and Networking (MobiCom)*, tutorial paper. https://doi.org/g4cp

- Han*, J., Xia*, T., **Spathis, D.**, Bondareva, E., Brown, C., Chauhan, J., Dang, T., Grammenos, A., Hasthanasombat, A., Cicuta, P., Floto, A., & Mascolo, C. (2021). Sounds of COVID-19: exploring realistic performance of audio-based digital testing. *Nature Digital Medicine (npj Digit. Med.), in press*. *equal contribution https://arxiv.org/abs/2106.15523

- Searle, B., **Spathis, D.**, Constantinides, M., Quercia, D., & Mascolo, C. (2021). Anticipatory Detection of Compulsive Body-focused Repetitive Behaviors with Wearables. *ACM International Conference on Mobile Human-Computer Interaction (MobileHCI)*, Toulouse & Virtual, France. https://doi.org/gzxx

- Han, J., Brown*, C., Chauhan*, J., Grammenos*, A., Hasthanasombat*, A., **Spathis*, D.**, Xia*, T., Cicuta, P., & Mascolo, C. (2021). Exploring automatic

COVID19 diagnosis via voice and symptoms from crowdsourced data. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto & Virtual, Canada. *equal contribution https://doi.org/gc25

- Tang, C., Pozuelo*, I., **Spathis*, D.**, & Mascolo, C. (2021). SelfHAR: Improving Human Activity Recognition through Self-training with Unlabeled Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT/Ubicomp), 5 (1)*. *equal contribution https://doi.org/f6tv

- Schuller, B., Batliner, A., Bergler, C., Mascolo, C., Han, H., Lefter, I., Kaya, H., Amiriparian, S., Baird, A., Stappen, L., Ottl, S., Gerczuk, M., Tzirakis, P., Brown, C., Chauhan, J., Grammenos, A., Hasthanasombat, A., **Spathis, D.**, Xia, T., Cicuta, P., Rothkrantz, L., Zwerts, J., Treep, J., & Kaandorp K. (2021). The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. *Proceedings of the Conference of the International Speech Communication Association (Interspeech)*, Brno & Virtual, Czech Republic (to appear). https://arxiv.org/abs/2102.13468

- Pozuelo, I., **Spathis, D.**, Gifford-Moore, J., Morley, J., & Cowls, J. (2021). Digital Phenotyping and Sensitive Health Data: Implications for Data Governance. *Journal of the American Medical Informatics Association (JAMIA)*. https://doi.org/fsxq

- Pozuelo, I., **Spathis, D.**, Clifton, E., & Mascolo, C. (2021). Wearables, smartphones and artificial intelligence for digital phenotyping and health. *Digital Health, Elsevier*, pp. 33–54 (ISBN: 9780128200773). https://doi.org/fpfv

- Greenberg, D., Wride, S., Snowden, D., **Spathis, D.**, Potter, J., & Rentfrow, J. (2021). Universals and variations in musical preferences: A study of preferential reactions to Western music in 350,000 people across 53 countries, *Journal of Personality and Social Psychology*, in press.

- Tang, C., Pozuelo, I., **Spathis, D.**, & Mascolo, C. (2020). Exploring Contrastive Learning in Human Activity Recognition for Healthcare. *Advances in Neural Information Processing Systems (NeurIPS-W), Machine Learning for Mobile Health workshop*, Virtual event, Canada. https://arxiv.org/abs/2011.11542

- Taquet, M., **Spathis, D.**, Rentfrow, J., Goodwing, G., & Mascolo, C. (2020) Improving the definition of depressed mood with digital phenotyping. *SSRN preprint*. https://doi.org/gmf5

- Pozuelo*, I., Posa*, M., **Spathis, D**., Westgate, K., Wareham, N., Mascolo, N., Brage, S., & Palloti*, J. (2020) Detecting sleep in free-living conditions without sleepdiaries: a device-agnostic, wearable heart rate sensing approach. *medRxiv preprint*. *equal contribution https://doi.org/gmf7

- Brown*, C., Chauhan*, J., Grammenos*, A., Han*, J., Hasthanasombat*, A., **Spathis*, D**., Xia*, T., Cicuta, P., & Mascolo, C. (2020). Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, San Diego & Virtual, USA. *equal contribution http://doi.org/d683

- **Spathis, D**., Passalis, N., & Tefas, A. (2019). Interactive dimensionality reduction using similarity projections. *Knowledge-Based Systems, 165:77-91*. http://doi.org/cxbm

- **Spathis, D**., Passalis, N., & Tefas, A. (2018). Fast, Visual and Interactive Semi-supervised Dimensionality Reduction. *European Conference on Computer Vision (ECCV-W), Workshop on Compact and Efficient Feature Representation and Learning in Computer Vision*, Munich, Germany. http://doi.org/cz6d

Chapter 2

Background

If I have seen further it is by standing on the shoulders of Giants
–Isaac Newton

In the previous chapter, we highlighted the importance of developing new models for mobile and wearable health data. In this chapter, we delve into the data modalities, tasks, and the fundamentals of relevant machine learning techniques used throughout this thesis. We first provide an introduction to machine learning for mobile health (Section 2.1), and we proceed with the building blocks of neural networks (2.2), common training paradigms (2.3), as well as considerations specific to time-series modeling (2.4).

2.1 Machine learning for mobile health

Mobile health is the application and use of mobile devices to healthcare. When combined with predictive capabilities, it offers novel and scalable ways to track and diagnose diseases, while reducing costs of the broader health system. Mobile health applications have been successfully introduced to model a variety of health outcomes such as infectious diseases (Wood et al., 2019), HIV medication adherence (Rana et al., 2016), and asthma management (Chan et al., 2017). The first large-scale studies focused on associations between control and experimental groups while making sure that they address challenges such as reporting–selection bias and low retention rates. Only recently machine learning has started to be used in an end-to-end way in mobile health. For example, an image-based deep learning system was evaluated by Google for automatic diagnosis of skin conditions (Liu et al., 2020), and was later rolled out in a mobile app for end users (Peggy Bui, 2021).

However, a sentiment echoed by the –still nascent– health informatics research community is that the development of such technologies has progressed at a faster

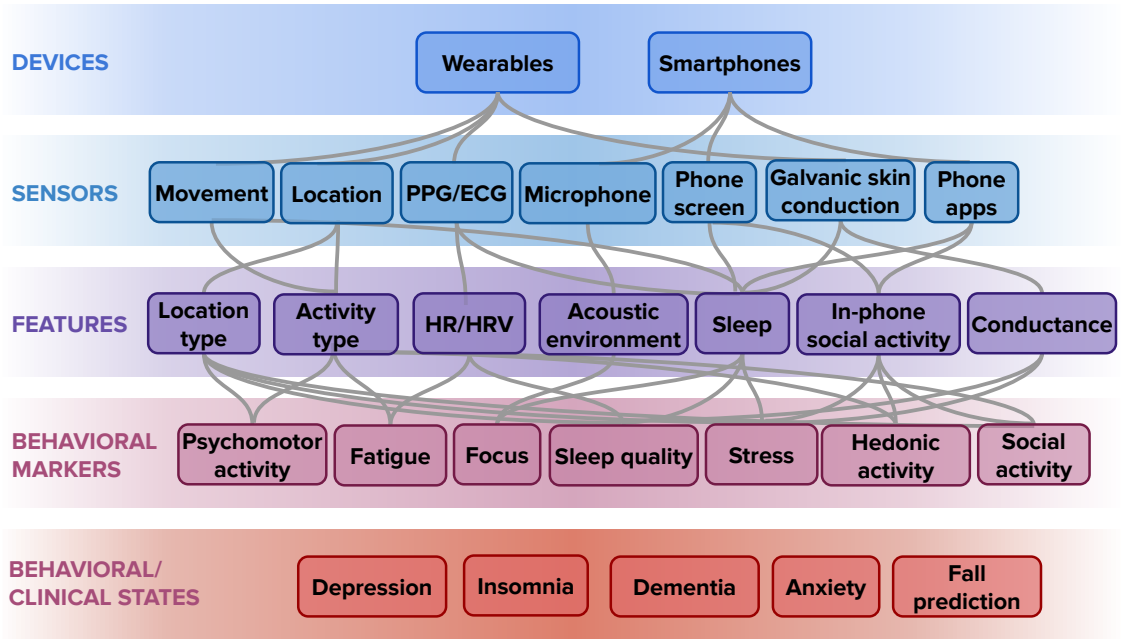


Figure 2.1: **Example of a hierarchical framework of wearable and mobile health**. The boxes at the top represent inputs to models. The boxes in between represent features and high-level behavioral markers along with outcomes. (PPG: Photoplethysmography, HRV: Heart rate variability). Figure inspired by (Mohr et al., 2017).

pace than the methods to evaluate and validate their efficacy. For example, commercial smartwatches are able to capture heart rate data accurately but fail in more complex metrics such as the *energy expenditure*, when compared to ground truth devices (Shcherbina et al., 2017). Also, context matters, especially considering that such devices can be used everywhere by everyone. For instance, skin tone or the type of activity (e.g. running versus walking) might affect the heart rate estimation (Bent et al., 2020). While most of these devices are increasingly using machine learning to estimate more high-level outcomes such as fitness or mood swings, we should ensure that they are being evaluated in diverse population cohorts.

Studies involving machine learning in mobile health usually fall into two categories; either proving the feasibility of the respective modality (sensors, images etc) to predict the respective outcomes (Wang et al., 2014), or, given a standardized dataset with a clearly defined outcome, trying to devise more accurate models (Aggarwal et al., 2019). As such, the level of sophistication in machine learning models has been usually commensurate with the dataset size (see Section 2.4 for an in-depth discussion of this

trade-off). Nevertheless, we argue for a *third way*; we collect large longitudinal datasets, identify bottlenecks of existing models when dealing with such datasets, and then propose bespoke models and pipelines that outperform other baselines.

Acknowledging that the term "mobile health" is quite broad, the scope of this Chapter (and this thesis) will be narrowed to devices with connected sensors and their associated mobile applications, excluding non-internet sensors such as thermometers, electronic health records, or large stationary equipment such as smart treadmills. We point the interested reader to this excellent review which covers all device categories (Marra et al., 2020). In all cases, we study data coming from smartphones and wearables as well as models which extract information from sensors in order to transform them to behavioral markers which eventually map to clinical states (see Fig. 2.1). To showcase the impact of mobile technologies, we focus on two broad areas which concern most people, that is, *mental* and *physical* health.

2.1.1 Mental health

Clinical outcomes and measurements. Mood and general mental wellbeing have been associated with several clinical outcomes. Self-reported sadness was found to be an indicator of depression (Cheng and Furnham, 2003), while self-reported happiness is linked to longevity (Veenhoven, 2008), personality traits (Ching et al., 2014; Geukes et al., 2017), and reduced mortality risk (Aichele et al., 2016). The experience sampling method (ESM), or ecological momentary assessment (EMA), –which involves asking participants to report their behaviors or environment on repeated occasions over time– has long been used as a mechanism to longitudinally assess the mental health of individuals by prompting them to report their mental state using questionnaires, traditionally administered using pen and paper, and also via the web. Psychologists have used different tools or scales to measure mood. These include the Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988), a self-report questionnaire of two 10-item scales that measures both positive and negative affect; and the Affect Grid scale (Russell et al., 1989), a 2-dimensional grid, where the x-axis indicates the feeling in terms of its positiveness or negativeness and the y-axis indicates its intensity. Independently of the scale used, timely and accurate mood reporting is important to anticipate clinical outcomes. To this end, smartphones and wearable devices have enabled timely delivery of experience sampling (Csikszent and Larson, 2014), allowing a near real-time detection of clinical outcomes and relapses.

Mobile devices brought increased reach. The penetration of mobile devices has introduced scale: many more individuals can now be reached and assessed. For example, in a hospital environment, mobile experience-sampling enabled the collection of 11,381

survey responses over a 12-month period from 304 physicians and nurses, completed with minimal initial training (Tejani et al., 2010). Mobile sensors enable researchers to collect not only the explicit reports of the participants, but also the *context* in which these answers were provided. Indeed, a survey of 110 ESM papers concluded that a total of 70 studies (63.6%) passively or actively collected sensor data from the participants' study device (Van Berkel et al., 2018). For instance, *StudentLife* (Wang et al., 2014) combined sensing and self-reports to assess the impact of student workload on stress, whereas *Snapshot* (Taylor et al., 2017) tracked their mood and sleep. Others focused on detecting depression by tracking medication, sleep patterns and actions (Suhara et al., 2017), location (Canzian and Musolesi, 2015; Palmius et al., 2016) or keypress acceleration (Cao et al., 2017). On a larger scale, *Utsureko* (Suhara et al., 2017) and *EmotionSense* (Servia-Rodríguez et al., 2017), two independent smartphone applications for mood monitoring through self-reports, were used by more than 24,000 and 17,000 users, respectively.

Towards larger datasets. However, most existing works suffer firstly from limited sample size, both in terms of number and diversity, which hampered them from drawing robust conclusions, and secondly from limited duration of the studies. For instance, in the *MoodScope* study (LiKamWa et al., 2013) 32 people were monitored for 2 months; in *StudentLife* (Wang et al., 2014), 48 students were tracked for 10 weeks, whereas in *Snapshot* (Taylor et al., 2017), probably the biggest general published study about mood monitoring using mobile devices, 206 students were tracked for 1 month. In contrast, this thesis employs the *EmotionSense* dataset (Servia-Rodríguez et al., 2017), which tracked tens of thousands of participants in the wild for more than 3 years, by collecting ground truth through self-reports as well as passive sensor data. Putting aside the limitations of the sample size, perhaps the most closely related work to ours is the *Snapshot* (Sano, 2016) study. This study investigated how daily behavior gathered through passive sensing data influence sleep, stress, mood, and other wellbeing-related factors. Multiple papers focused on different aspects of the collected dataset, such as personalization with multi-task learning to predict tomorrow's mood, stress, and health (Taylor et al., 2017), and the prediction of happy/sad moods based on sleep history (Sano et al., 2015).

The *EmotionSense* dataset. This thesis uses the *EmotionSense* dataset (Servia-Rodríguez et al., 2017), a dataset that contains sensor and self-reported data collected with a mobile phone application for Android designed to study subjective well-being and behavior. From February 2013 until October 2016[1], this application collected 735,778 self-report datapoints from 17,251 users, through surveys presented on the

[1]We use a longer time-frame of collected data than that of (Servia-Rodríguez et al., 2017). Their ending timestamp was in January 2016.

phone via experience sampling, and behavioral data from physical and software sensors in the phone (accelerometer, microphone, location, text messages, phone calls, etc.). The participants signed a consent form that restricts the use of the data to the University of Cambridge researchers, in accordance with Institutional Review Board (IRB) protocol. For this analysis, we consider self-reported mood collected graphically using the Affect Grid (Russell et al., 1989), profile-related surveys, as well as sensed data collected with the accelerometer and microphone sensors. Twice per day, between 8am and 10pm after a time interval of at least 120 minutes, participants received a notification asking them to report their mood in the affect grid. Meanwhile, sensed data was collected passively in the background at different time-points during the day depending on the different versions of the study. At different stages of the application, participants were requested to complete profile-related questionnaires covering a broad range of topics including: demographics, personality, gratitude, health, sociability, job satisfaction, life aspirations and connectedness, where the questions were answered using Likert scales.

Machine learning models. Regarding machine learning models, the majority of related literature has applied some kind of supervised learning, such as Logistic Regression or Support Vector Machines, which cannot capture non-linear combinations of features. For a more extensive view of the field, we point the interested reader to this comprehensive review on ML for mental health (Thieme et al., 2020). Some recent works employ RNNs (Suhara et al., 2017), feedforward layers (Mikelsons et al., 2017), multi-task learning (Taylor et al., 2017), and autoencoders to fill in missing sensor data (Jaques et al., 2017), or to learn better representations (Li and Sano, 2020; Liu et al., 2020). Further, binary prediction is quite common in the mood prediction literature, where mood is simplified to a binary state (Taylor et al., 2017; Servia-Rodríguez et al., 2017), so that for instance, extreme depression is binned in the same class as moderate unhappiness. Since neutral mood might be uninformative and make the predictions harder, authors often omit the middle-scoring 40-60% of reports. Instead, we explore fine-grained mood prediction through regression and clustering, as well as novel multi-dimensional formulations of the Affect Grid through multi-task learning.

2.1.2 Physical health

Measuring physical activity at scale. Large scale studies of physical activity and well-being leveraging mobile devices' built-in accelerometers have shown promise as global physical activity surveillance tools, demonstrating, for example, inequalities across different countries after analysing data from over 700, 000 people (Althoff et al., 2017). Another analysis of exercise patterns in a global social network of 1,1 million

runners, demonstrated that exercise is "contagious", whose effect depends on gender and relative levels of activity (Aral and Nicolaides, 2017). 10 million users of a weight monitoring app were used to show that people are more likely to lose weight when they had more friends of the opposite sex (Wang et al., 2017). Weight loss was the subject of other studies of over 1 million participants (Serrano et al., 2017), which showed that power users demonstrated the greatest weight loss. The relationship between physical activity and cardiovascular disease was studied for 50,000 people in (McConnell et al., 2017), finding that lower overall activity but more frequent transitions between active and inactive periods was associated with similar cardiovascular disease to higher overall activity but with fewer transitions. Mobile and wearable sensors allow for continuous and ubiquitous monitoring of an individual's physical activity profiles, which, when combined with cardio-respiratory information, provide valuable insights into that individuals' health and fitness status (Mandsager et al., 2018). Hence, the possibility of measuring individuals' physiological characteristics in free-living conditions is of great interest for research, clinical and commercial applications. In particular, 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 (Jones and Carter, 2000)), thus, leveraging these two signals concurrently likely produces better representations than either signal taken in isolation. For instance, 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 (Savonen et al., 2006). In healthy individuals, HR responses to activity are defined by an increase in HR that is concurrent to the increasing intensity of the activity (Ellestad and Wan, 1975).

Challenges in modeling wearable data. 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 (Blalock and Guttag, 2016). 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 (Doherty et al., 2017; Menai et al., 2017)[2]. However, since wearables are used in daily, unconstrained environments, activities such as drinking coffee or alcohol, as well as stress, may confound simple heuristics.

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

Fitness as a key predictor of well-being. A concept that is particularly central to physical activity is that of cardio-respiratory fitness (CRF), an important modifiable marker of cardiovascular health embodied by a strong inverse relationship with the incidence of cardiovascular disease (CVD), type 2 diabetes, cancer, mortality and other adverse health outcomes (Lynch et al., 1996; Lakka et al., 1994; Myers et al., 2002; Ekelund et al., 1988; Schmid and Leitzmann, 2015; Schuch et al., 2016; Blair et al., 1989; Laukkanen et al., 2004; Mandsager et al., 2018). Clinical clinical evidence shows that CRF is not only potentially a stronger predictor of mortality than well-established risk factors like hypertension, type 2 diabetes, high cholesterol or smoking, but that using CRF to complement these traditional risk factors significantly improves the precision of risk prediction models for adverse CVD health outcomes (Ross et al., 2016; Myers et al., 2002; Kokkinos et al., 2013; Lloyd-Jones et al., 2010). Beyond its implications in medicine, CRF is frequently used in sports as indicator of endurance capacity, having strong predictive value for other sport-related performance traits (Ross et al., 2016).

Challenges in collecting fitness ground truth. The *gold-standard* measure of CRF is the maximal oxygen uptake (VO_2max), which measures the maximal rate at which an individual can consume oxygen during exercise. VO_2max is assessed through an exercise test to exhaustion while respiratory gas exchange is measured, with the assessment only deemed a true maximal result if several test criteria are met. These criteria include leveling-off of oxygen uptake and heart rate (HR) and the surpassing of thresholds for the respiratory exchange ratio. This type of assessment requires trained staff and expensive laboratory settings with specialized equipment and often test criteria for maximal effort are not met (Swain et al., 2014; Davis, 1995). Given these logistical constraints and the inherent risk of maximal exercise testing, scalability of fitness assessment in large populations has been limited, meaning relatively little is known about population levels of fitness, or their possible changes over time.

Scaling fitness prediction. Despite some promising studies which attempt to infer VO_2max from data collected during free-living conditions, these mostly stem from small-scale cohorts with less than 50 participants and use contextual data from treadmill activity, which again limits their application in real-world contexts (Altini et al., 2016). In this thesis, we employ data from the *Fenland Study*, the largest study of its kind, following more than 10,000 participants for a week and almost a decade later to assess the change of fitness. We use purely free-living data to predict VO_2max.

The *Fenland* dataset. The *Fenland* study is a prospective cohort study that includes 12, 435 men and women who are between the ages of 35 and 65 (O'Connor et al., 2015). 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 National Research Ethics Service - Cambridge East Research Ethics Committee (IRAS ID 138617). 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 hours per day 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 *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) (Brage et al., 2004). The *Fenland* study has two distinct phases. Phase I, during which baseline data was collected from 12,435 participants, took place between 2005 and 2015. Phase II was launched in 2014 and involved repeating the measurements collected during Phase I, alongside the collection of new measures. All participants who had consented to being re-contacted after their Phase I assessment were invited to participate in Phase II. At least four years must have elapsed between visits. As a result of this stipulation, recruitment to Phase II is ongoing. A subset of 2,675 of the study participants returned for the second phase of the study, after a median (interquartile range) of 6 (5-8) years, and underwent a similar set of tests and protocols, including wearing the combined heart rate and movement sensing for 6 days.

Modeling wearable signals with machine learning. Machine learning models have been only recently applied to this task. To approximate VO_2max without the need for a dynamic test, non-exercise models aim to provide a viable alternative to CRF assessment for widespread use in many healthcare settings. These are usually traditional regression models and incorporate variables like sex, age, body mass index (BMI), resting heart rate (RHR) and self-reported physical activity to infer VO_2max (Cao et al., 2010; Jurca et al., 2005). However, the validity of such estimation is still much lower than what can be achieved with dynamic exercise testing (Gonzales et al., 2020; Nes et al., 2011). Wearable devices, such as activity trackers and smartwatches, increasingly provide opportunities for non-intrusive objective monitoring of biological signals such as heart rate and movement during free-living, potentially enabling more precise prediction of VO_2max without the need to conduct a specific exercise test (Plasqui and Westerterp, 2005).

Deep learning for physical activity and vitals. Recent advances in deep learning architectures for sequential modeling based upon wearable and mobile sensing

have been used for health predictions and recommendations (Ballinger et al., 2018; Schwab and Karlen, 2019). 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 (Ni et al., 2019). This approach is helpful but requires prior segmentation of activities, which can be a constraint when applying these techniques in free-living, unconstrained conditions. Recent work using self-supervised learning has shown promise in the same data modalities such as ECG data (Sarkar and Etemad, 2019; Hallgrímsson et al., 2018; Kiyasseh et al., 2020). In the following sections we expand on these new neural network training paradigms and discuss how we build upon them towards more accurate fitness prediction as well as generalized models which can predict multiple physiological outcomes.

2.2 Deep neural networks

The history of neural networks dates back to the 1950s with the invention of the *perceptron* (Rosenblatt, 1958), which paved the way for today's modern Deep Neural Networks (DNNs) (LeCun et al., 2015). The goal of a neural network is to approximate some function f, so that, for example, a classifier $\mathbf{y} = f(\mathbf{x};\boldsymbol{\theta})$ maps the input $\mathbf{x}$ to a category or real value $\mathbf{y}$, by learning the optimal parameters $\boldsymbol{\theta}$ that best approximate this function. They are called *networks* due to the fact that they are composed of many different functions. For example, four functions –or *layers*– $f^{(1)}$, $f^{(2)}$, $f^{(3)}$, and $f^{(4)}$, can be connected in a chain to form a computational graph:

$$f(\mathbf{x}) = f^{(4)}(f^{(3)}(f^{(2)}(f^{(1)}(\mathbf{x})))) \tag{2.1}$$

where stacking several layers defines the *depth* of the model, hence the widely adopted term *deep learning*.

More formally, considering a simplified case of a two-layer network and an output unit, we define it as follows:

$$\begin{aligned} \mathbf{z}_1 &= g(\mathbf{W}_1^\top \mathbf{x} + \mathbf{b}_1) \\ \mathbf{z}_2 &= g(\mathbf{W}_2^\top \mathbf{z}_1 + \mathbf{b}_2) \\ \mathbf{out} &= \sigma(\mathbf{W}_3^\top \mathbf{z}_2 + \mathbf{b}_3) \end{aligned} \tag{2.2}$$

where $\mathbf{z}$ are intermediate layers with trainable weights $\mathbf{W}$ and biases $\mathbf{b}$, and $\mathbf{out}$ is the output of the model. The result of each layer is further processed by an activation function, which, like g, is usually a Rectified Linear Unit (ReLU($\mathbf{z}$) = $max(0, \mathbf{z})$). The

intuition behind this function is that it allows the network to learn non-linear patterns of the features. Last, the output layer activation σ can be a *softmax* function $\left(\frac{e^{\mathbf{z}_i}}{\sum_{j=1} e^{\mathbf{z}_j}}\right)$, to transform the output to a probability distribution over predicted classes (classification). In regression problems, the raw numerical output can be the final prediction, and usually no further activations are applied.

To train the model, we use *backpropagation* (Linnainmaa, 1970), which defines the error and in turn compares the expected output to the ground truth $\mathbf{t}$:

$$\mathbf{E} = \frac{1}{2} \|(\mathbf{out} - \mathbf{t})\|_2^2 \tag{2.3}$$

This error is calculated in the forward pass of the neural network. Then, it is *propagated* across all the intermediate layers, which allows the model to learn from its mistakes and repeat the training process by adjusting the weights. Formally, given a single weight $\mathbf{W}_i$ and the previous layer output $\mathbf{z}_i$, we define the backward pass as follows:

$$\frac{\partial \mathbf{E}}{\partial \mathbf{W}_i} = \frac{\partial \mathbf{E}}{\partial \mathbf{z}_i} \frac{\partial \mathbf{z}_i}{\partial \mathbf{W}_i} \tag{2.4}$$

After computing the gradients, the network updates the weights according to the following equation:

$$\mathbf{W}_i = \mathbf{W}_i - \alpha \frac{\partial \mathbf{E}}{\partial \mathbf{W}_i} \tag{2.5}$$

where α is a scalar called *learning rate* and its value is decided by the optimization algorithm of choice. For an in-depth view of deep learning, we recommend the excellent Deep Learning book (Goodfellow et al., 2016). Next, we discuss the shortcomings of perceptron (feedforward) layers when processing data with dependencies (e.g. sensor timeseries) and present modern approaches which handle these challenges.

2.2.1 Recurrent neural networks

While feedforward layers can easily handle two-dimensional inputs $\mathbf{x} = (\mathbf{x}_1,...,\mathbf{x}_N) \in \mathbb{R}^{N \times F}$, where N are the samples and F the features, data from complex systems is usually high-dimensional with internal dependencies. For example, sequential data is commonly represented as three-dimensional tensors $\mathbf{x} = (\mathbf{x}_1,...,\mathbf{x}_N) \in \mathbb{R}^{N \times T \times F}$, where T are the timesteps. In other words, every sample includes both a timeseries and

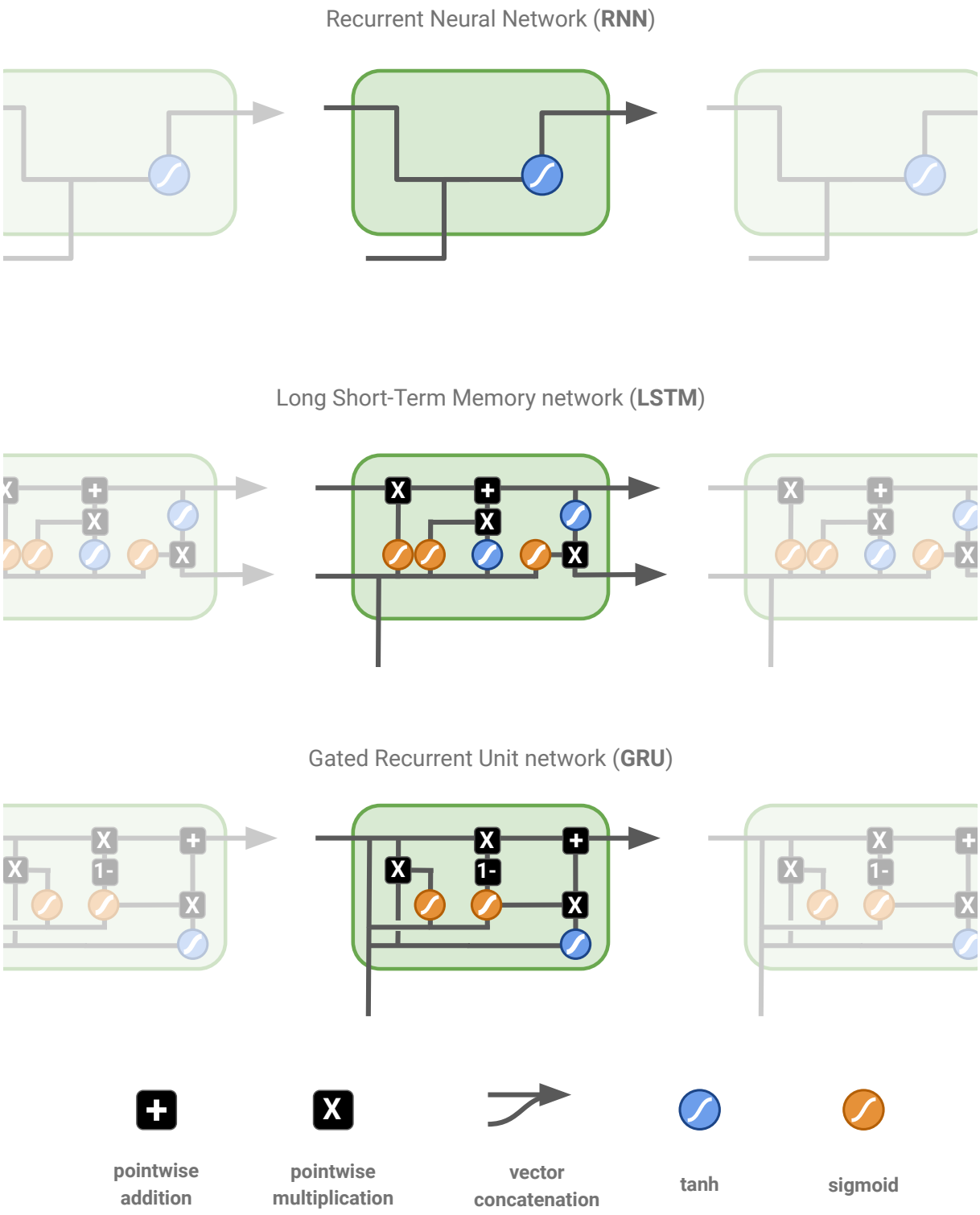


Figure 2.2: **Variants of recurrent neural networks**. Conceptual illustration of the various components and differences between RNNs, LSTMs, and GRUs. Illustration inspired by (Olah, 2015).

its associated features [3]. In practical terms, a feedforward layer can *still* ingest such tensors, however, it will not learn any sequence-aware patterns because the tensor should be flattened ($T \times F$) prior to the initial dot product with the weights.

As Christopher Olah put it so eloquently *"Humans don't start their thinking from scratch every second. You don't throw everything away and start thinking from scratch again. Your thoughts have persistence* (Olah, 2015)". Feedforward layers cannot do this. Therefore, new layers have been proposed which *share* the same weights across several time steps and, as a result, are able to explicitly leverage and process sequential data. In this thesis, we mainly use variants of *Recurrent Neural Networks* (RNNs) and *Convolutional Neural Networks* (CNNs).

An RNN (Rumelhart et al., 1986) is a specialized layer for processing a sequence of values. In Eq. 2.1, a classical computational graph just passes through information across layers, one after another. On the other hand, RNNs introduce extra feedback loops which allow the models to learn different weights for steps of an input sequence. More formally, the RNN has three main weights: the input to hidden connections is parametrized by a weight matrix $\mathbf{U}$, hidden-to-hidden recurrent connections are parametrized by $\mathbf{W}$, and hidden-to-output connections are parametrized by $\mathbf{V}$. The internal dynamics for each timestep t of the RNN are defined as follows:

$$\begin{aligned} \mathbf{a}^{(t)} &= \mathbf{b} + \mathbf{W}\mathbf{h}^{(t-1)} + \mathbf{U}\mathbf{x}^{(t)} \\ \mathbf{h}^{(t)} &= tanh(\mathbf{a}^{(t)}) \\ \mathbf{o}^{(t)} &= \mathbf{c} + \mathbf{V}\mathbf{h}^{(t)} \\ \hat{\mathbf{y}}^{(t)} &= softmax(\mathbf{o}^{(t)}) \end{aligned} \tag{2.6}$$

where $\mathbf{h}$ is the hidden state, $\mathbf{b}$ and $\mathbf{c}$ are biases, and $\hat{\mathbf{y}}$ the output probabilities. While RNNs have proved particularly effective in text prediction tasks (Nallapati et al., 2016), scalable training is problematic because the gradients either *explode* or *vanish* at each time step. As a result, although the main premise is to learn long-term dependencies, evidence shows that it is difficult to learn patterns in long sequences (Bengio et al., 1994).

To correct for that, an idea was to add explicit memory to RNNs. The first proposal of this kind is the *Long Short-Term Memory* (LSTM) (Hochreiter and Schmidhuber, 1997). These networks introduce special *gates* that filter the information as it flows across the layer and allow the LSTM to keep or forget information. The core concept is the *cell state* which acts as a conveyor belt, or the memory of the network. As the cell state is

[3] Sometimes, the third component of a 3D tensor which hosts multiple parallel timeseries is called a *channel*.

progressing, the gates, which are different neural networks, decide which information is allowed on the cell state. In particular, the *forget gate* determines what information should be stored by applying a sigmoid function. Then, the *input gate* updates the cell state through a combination of sigmoid and tanh functions. The cell state, is now updated and the *output gate* decides what the next hidden state should be, through another pair of sigmoid and tanh functions. LSTMs have achieved state of art results in many sequence problems, with a particular success in speech recognition (Graves et al., 2013).

However, at this point it is natural to wonder which parts of the LSTM are actually necessary. Recent variants of the LSTM include the *Gated Recurrent Units* (GRUs) (Cho et al., 2014), which use the hidden state to transfer information, doing without the cell state of the LSTM. In particular, a GRU has only two gates, a *reset gate* which decides how much past information to retain, and an *update gate* which combines the functionality of the forget and input gates of the LSTM. A conceptual illustration of the variants of recurrent networks can be found in Fig. 2.2, where one can observe the increasing complexity of LSTMs and GRUs over the RNN.

Another useful property of RNNs/LSTMs is the ability to forecast multiple timesteps. Hidden Markov Models (HMMs), autoregressive models and regression algorithms have been traditionally applied to sequence prediction. HMMs and autoregressive models operate by default on single sequences, being unable to learn patterns from several users. Traditional feature-based ML algorithms such as linear regression, random forests or support vector regressors, can solely predict one scalar value. They do not support an extended forecast horizon without feeding through the previous prediction as its new input (Venkatraman et al., 2015), which unavoidably introduces compounding errors that skew the input distribution for future prediction steps. RNNs have become increasingly popular in modeling sequential, high-dimensional, non-linear data by incorporating encoder-decoder architectures (Figure 2.4). Recent RNN models (named sequence-to-sequence or *seq2seq*) can map an input sequence to an output sequence of any arbitrary length, making RNNs the state-of-the-art in Natural Language Processing for machine translation and speech processing (Sutskever et al., 2014) since they can map, for example, a phrase in French to a phrase of different length in English.

2.2.2 Convolutional neural networks

An effectively simpler approach to timeseries modeling is to repurpose the convolutional operations that have been successfully applied to image processing (Krizhevsky et al., 2012). Traditionally, the image processing community would use the convolution

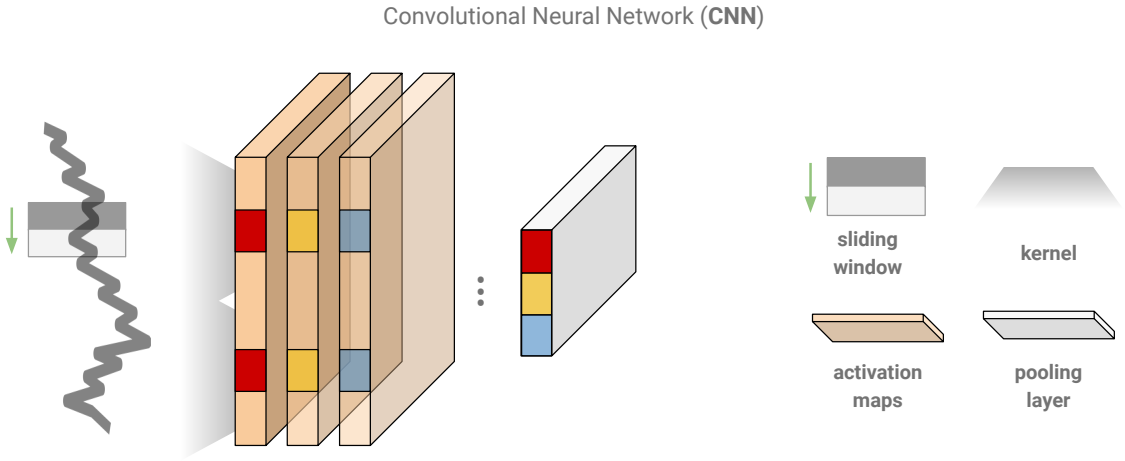


Figure 2.3: **Convolutional neural networks**. Conceptual illustration of the components of a convolutional neural network applied to timeseries data.

operation (or *kernels*) for tasks such as blurring or sharpening. In its simplest form, a convolution is done by multiplying a pixel and its immediate neighborhood with a matrix. The same idea is applied to CNNs, where every subsequent layer learns feature maps of increasing or decreasing granularity by scanning the 2D space of an input image.

Until recently, sequence modeling was synonymous with RNNs. However, empirical results suggest that 1D CNNs are equally good –or better– in a set of diverse tasks (Bai et al., 2018). This idea goes back to the 1980s where *time-delay networks* first popularized the notion of shift-invariance in time: the network would discover temporal relationships which will not be confounded by temporal shifts of the input (Waibel et al., 1989). In addition to better performance, CNNs are faster to train, which justifies the resurgent interest in the ML community for sequential tasks.

More formally, given an input dataset $\mathbf{x} = (\mathbf{x}_1,...,\mathbf{x}_N) \in \mathbb{R}^{N\times T\times F}$, 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

$$C(s) = (\mathbf{x} * f)(s) = \sum_{i=0}^{k-1} f(i) \cdot \mathbf{x}_{s-i} \tag{2.7}$$

where k is the filter size, $s-i$ records the convolution step and $*$ denotes the convolution operator. We note that the 1D window learns patterns across all the parallel features of the 3D input tensor $\mathbf{x}$. Often, after the convolutional layers the dimensionality of the resulting feature maps is further reduced with a *pooling* layer, which either

aggregates the maps with a max or averaging operation (same dimensions but fewer parameters), or –more aggressively– performs *global* pooling (reduced dimension)[4]. A conceptual illustration of a CNN applied to timeseries data can be found in Fig. 2.3, where one can observe how the learned CNN features are further summarized by pooling operations.

1-dimensional CNNs have been applied to various fields with remarkable success. For example, some of the most widely-used audio synthesis and language modeling systems are based on CNNs (Van Den Oord et al., 2016; Dauphin et al., 2017).

2.2.3 Other layers

Apart from RNNs and CNNs, other layers have been proposed over the past years to model sequence data with neural networks. Most notably, *attention* layers leverage dot products of the input data in combination with simple feed-forward networks (Bahdanau et al., 2014). The intuition is based on how we pay visual attention to different regions of an image or correlate words in one sentence while reading. Pure attention architectures –like the *Transformer* or *BERT* (Vaswani et al., 2017; Devlin et al., 2019)– have dominated language tasks, however, in the case of high-dimensional signals, the literature has not "converged" to attention-only networks yet. Instead, it is common to mix CNN/RNNs with further attention layers (Hao and Cao, 2020). This might be due to the structural differences between language and time-series data; memory-aware layers may explicitly capture the temporal patterns of continuous data, whereas discreet data (text) could be favored by attention.

Another approach tailored to irregularly sampled timeseries was to supercharge RNNs with properties of ordinary differential equations (ODEs). ODE-RNNs (Rubanova et al., 2019) achieve strong results in datasets with missing timesteps, when compared to simpler methods like GRU-D which tweak the RNN's internal state to account for the time delta between the datapoints (Che et al., 2018).

On the other hand, we see that sometimes the data sampling method is *more important* than the layer of choice. For instance, it is worthwhile to mention *MLP-Mixer* (Tolstikhin et al., 2021), a recent method that achieves state of art results in vision with simple feed-forward layers (no CNN/RNN/Attention), by splitting the input images into patches. This thesis advocates for and presents some new data alignment methods towards better representation learning. In the next section, we discuss some notable training paradigms that are gaining momentum towards this direction, moving beyond

[4]Considering that neural networks can be seen as *LEGO™ blocks* whereby every layer has to match the dimensions of the next one, the output vector defines the intermediate operations. The data has to flow from the high-dimensional input to the –usually simpler– output, and hence different layers such as *Pooling* or *Flattening* are used to achieve that.

traditional supervised learning.

2.3 Training paradigms

After introducing the main building blocks of neural networks, we will now discuss some emerging training paradigms which are inspired by real-world tasks and their challenges. For example, humans perform thousands of tasks naturally, but our models are still painfully single-purpose. What does it take to bridge this gap? Also, infants are never explicitly taught physics or grammar, but they still learn to speak or the effects of gravity through observation and interaction with their surroundings (Vallabha et al., 2007). Likewise, can we apply this paradigm to machine learning? For instance, we never start learning something new from scratch, but we build upon basic abstractions. Our models can learn to re-use and fine-tune, but what are their limits?

As a means of answering these fundamental questions, this thesis is employing *multi-task learning*, *multimodal learning*, *self-supervision*, and *transfer learning*.

2.3.1 Multi-task learning

Multi-task learning is a training method in which a model learns to predict simultaneously two or more similar tasks. For example, an architecture with shared layers and separate outputs might enable to perform –with a single model– robotic grasping, pushing, and poking (Pinto and Gupta, 2017). Multi-tasking has been used to reduce overfitting on the main task (with *auxiliary targets*), produce better data representations, and in general to improve accuracy in neural networks (Ruder, 2017). Specifically in deep neural networks, this multi-target setup forces the shared weights of the network to optimize all tasks and consequently learn internal representations that draw from multiple outcomes.

More formally, let $\mathcal{T}$ be a set of learning tasks for a dataset $\{\mathbf{x},\mathbf{y}\}$. The model tends to share some weights θ_g on the lower layers, where it learns generic patterns that are useful to all tasks. As we move to the last layers, the model is split to various "forks" which host task-specific weights θ_τ (see Fig. 2.4). In particular, the training optimizes a joint multi-objective loss where each task contributes to the weighted sum as follows:

$$\mathcal{L}_{joint} = \sum_{\tau=0}^{\mathcal{T}} \omega_\tau \times \mathcal{L}_\tau \tag{2.8}$$

where $\mathcal{L}_\tau$ denotes the individual-task loss and ω_τ the task weights. Despite this joint objective formulation and the parameter sharing, the tasks should present some similar characteristics. Besides, the input data should be general enough so as to

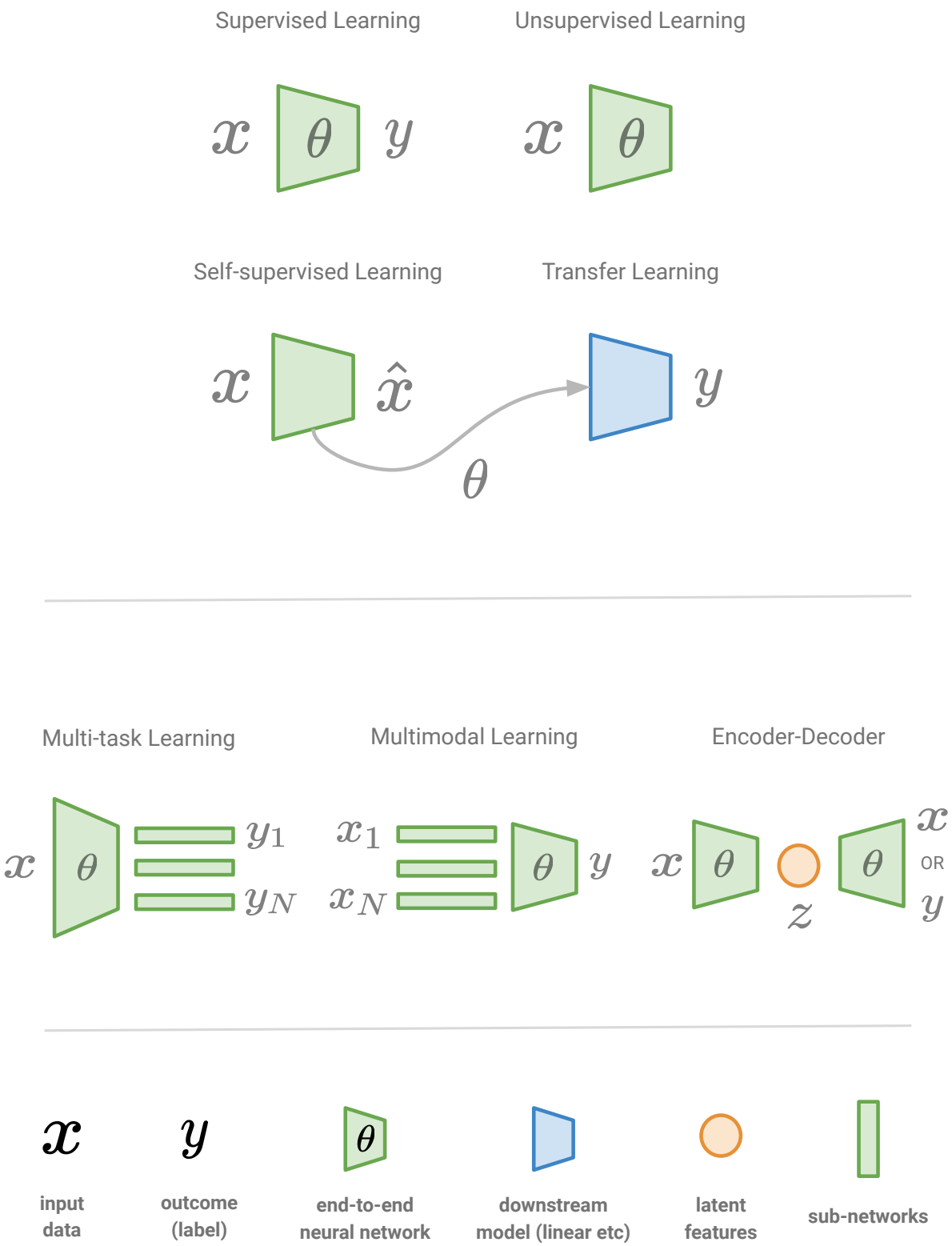


Figure 2.4: **Training paradigms**. Illustration highlighting the differences across supervised, unsupervised, and self-supervised learning, along with emerging neural network architectures applied to multimodal data.

generalize to many tasks; otherwise the singular tasks would dominate the loss. We point the interested reader to this survey which covers many recent architectures and applications of multi-tasking (Crawshaw, 2020).

Multi-task learning has been successfully applied to high-impact systems which range from single models that caption images and translate text (Kaiser et al., 2017), to self-driving cars which perform joint depth estimation and image segmentation (Kendall et al., 2018). In this thesis, we use multitask learning in conjunction with encoder-decoder models in Chapter 4.

2.3.2 Multimodal learning

While multi-task learning aims to better leverage the outputs or outcomes, *multimodal learning* involves relating information from multiple inputs and sources. An illustrative example comes from speech recognition: a visual */ga/* with a voiced */ba/* is perceived as */da/* by most people. In other words, the so-called McGurk effect (Ngiam et al., 2011) showed that people tend to integrate audio-visual information in order to decode speech, and as such, the visual modality provides information which helps to disambiguate between similar acoustics.

Machine learning at its core is about finding better representations of data. When dynamic data comes from different sources (e.g., a motion sensor and an ECG), it may present non-linear correlations where it is not straightforward to match and relate raw time-series collected with different devices or sampling rates. On top of that, extra metadata encoded as static features (e.g., sex or height) have to be processed in isolation. Multimodal representations usually involve a single end-to-end network with individual sub-networks for every modality which are later merged in a joint representation (see Fig. 2.4). Similar to parameter sharing in multi-tasking, here we decide between *early* (or *data-level*) and *late fusion*, which defines the stage in which the latent features are merged. In general, late-fusion strategies are more intuitive, particularly when the modalities vary in terms of sampling and dimensionality and often result in better performance (Ramachandram and Taylor, 2017).

Nonetheless, reiterating the *Lego* analogy of neural network layers, a third category became possible with modern models, that of *intermediate fusion*. The latest models attempt to first transform all modalities into representations, and then it becomes possible to fuse different representations into a single hidden layer (or shared representation). A simple –yet effective– strategy is to concatenate all resulting representations, with increasingly more sophisticated ideas investigating loss functions which enforce inter- and intramodality correlations (Wang et al., 2015).

Multi-task and multimodal learning can be seen as two sides of the same coin, in

which we strive for enabling many capabilities and leveraging different sources with a *single* model. In this thesis, we use multimodal models in Chapters 3, 5, and 6.

2.3.3 Self-supervised learning

Given a training task and enough labels, supervised learning can achieve good performance. This performance usually requires a decent number of manual labels which –in the best scenario– might become easier through crowdsourcing (cf. Imagenet), but in some cases is virtually impossible. For example, annotating wearable sensor timeseries for human activity recognition tasks *a posteriori* is not feasible without a video recording. On the other hand, considering the amount of unlabeled data (e.g. all the images on the internet or the entire Fitbit user base) is considerably more than is contained within a limited curated dataset, motivates ongoing research in this area. However, unsupervised learning is hard and until recently was less efficient than supervised learning (Chen et al., 2020).

A simple yet exciting recent idea was to obtain labels "for free" from the input data ($\mathbf{x}$) through various transformations, and then use conventional supervised objectives to predict them ($\hat{\mathbf{x}}$). The representations obtained this way would be meaningful for downstream tasks with limited labeled data and linear classifiers (see Fig. 2.4). This is coined as *self-supervised* (or predictive) learning due to learning the supervision directly from the data[5]. Even before this term, researchers would attempt to find *surrogate* tasks (Dosovitskiy et al., 2015) which would exploit unlabeled data. The most common tasks involve predicting distorted versions of the spatial characteristics of image data by means of rescaling (Dosovitskiy et al., 2015), rotating (Gidaris et al., 2018), patching (Doersch et al., 2015), shuffling (Noroozi and Favaro, 2016), colorization (Zhang et al., 2016), and inpainting missing parts (Pathak et al., 2016).

These pre-training tasks have achieved state-of-the-art results in computer vision (Lee et al., 2017; Jenni and Favaro, 2018) and natural language processing (Lan et al., 2020). However, someone would argue that devising these increasingly complex pre-training tasks resembles traditional feature engineering that neural networks promised to automate. Therefore, more generic methods switched focus from data transformations to the loss function level by offering elegant methods of implicit clustering between pseudo– positive and negative samples. Notably, *SimCLR* (Chen et al., 2020) achieved –for the first time– performance on par with supervised models, by proposing a training method for visual representations which maximizes agreement

[5]The terminology surrounding unsupervised and self-supervised learning is a bit blurry. In strict terms, unsupervised used to mean methods for principal component and cluster analysis (no labels). Beyond the supervision dichotomy, reinforcement learning offers an alternative formulation to intelligence through reward-based learning, which is however out of scope of this thesis.

between differently transformed views of the same sample via a *contrastive* cosine similarity loss in the latent space. More recently, *BYOL* claimed better results even without the negative pairs in its training objective through a similar two-network approach (Grill et al., 2020).

It is not clear, however, whether these methods can generalize beyond the image domain, with preliminary results showing that they underperform without domain-aware modifications (Tang et al., 2020; Kiyasseh et al., 2020). On the other hand, the closest conceptual modality to mobile and wearable timeseries is video, in which the relevant literature attempts to leverage the temporal –rather than spatial– information. Namely, approaches like the *arrow of time* (Wei et al., 2018) posit that low-level physics (like the smoke rising up) and high-level events (e.g. you cannot revert breaking a glass), produce more effective representations in downstream action classification tasks. This thesis, along with recent works (Jawed et al., 2020; Taghanaki and Etemad, 2020; Chen et al., 2021), argues that models which anticipate and forecast the future are more robust and generalizable. We use principles of time-aware self-supervision in Chapters 4 and 5.

2.3.4 Transfer learning

Transfer learning is the natural application of self-supervised learning. The term *transfer* describes a set of methods towards preserving and reusing previously acquired information, applied possibly to a slightly different domain. This stored information can further accelerate the training of a *downstream* task with usually limited training data. For the context of this thesis –and considering that it is a well-studied problem with traditional methods–, we focus on transferring the weights (also known as *representations* or *embeddings*) of deep neural networks. Modern transfer learning uses pre-trained networks –either supervised or self-supervised– as fixed feature extractors in linear downstream models (e.g. logistic regression) or further fine-tuning by *freezing* the backbone architecture and retraining the last layer. This has shown remarkable results in vision and language domains (Chen et al., 2020; Devlin et al., 2019).

More formally, given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$, the goal of transfer learning is to improve the predictive performance of target function $f_T(\cdot)$ in $\mathcal{D}_T$, leveraging the knowledge of the source domain $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$, or $\mathcal{T}_S \neq \mathcal{T}_T$. In the special case when the target and source tasks are identical, it is known as *transductive transfer*, whereas in *inductive transfer*, the target task differs from the source task, no matter if the source and target domains are the same or not. *Multi-task learning* can be seen a special case of inductive transfer, where labeled data is available in the source domain and source/target tasks

are learnt simultaneously. We point the interested reader to this comprehensive survey (Pan and Yang, 2009).

In this thesis, we explore transfer learning in conjunction with self-supervision in Chapter 5 and implicitly through multi-task learning in Chapter 4.

2.4 Modeling time-series and signals

We saw that most novel models in deep learning are evaluated with either image or text data. We will now discuss whether we should take into account specific considerations when training ML models with high-dimensional signals.

In the following sections, two *schools of thought* are presented, in the context of mobile and wearable health data. The first seeks for informative features that represent time-series through inventive feature extraction, while the second is based on the emerging power of representation learning to automatically extract features from *lightly* processed time-series.

2.4.1 Traditional feature-engineering modeling

Most commonly, sensor data coming from personal devices is transformed to *feature vectors* in order to be compatible with the majority of machine learning algorithms. A feature vector is a spreadsheet-like data structure where each row is a unique sample and each column a different feature or variable. However, the raw readings coming from e.g. an accelerometer are represented as multiple continuous sequences. Consequently, the next step after the data collection is to *summarize* each sensor in a couple or more independent variables through a *sliding window* operation. This laborious task is called feature-extraction and researchers try to come up with increasingly more complicated features that explain the respective label (see Fig. 2.5). For example, the *MoodExplorer* study (Zhang et al., 2018) extracted the mean, variance, and signal-to-noise ratios from the microphone sensor, while the *EmotionSense* study (Servia-Rodríguez et al., 2017) calculated the standard deviation of the magnitude of acceleration ($\sqrt{x^2 + y^2 + z^2}$) from the three axes (x, y, z) of the accelerometer.

Depending on the size of the datasets and the computing power, computing these features as a pre-processing step can be time-consuming as well as renders the process multi-step. Simple statistics like the mean, median, standard deviation and inter-quartile ranges might be easier to estimate but they may not capture informative features of noisy signals. On the other hand, higher-order statistics and transformations like the kurtosis, skewness, stationarity, least squares slope, autocorrelation, Fourier transforms, and entropy, provide more expressive metrics that reflect real-world phe-

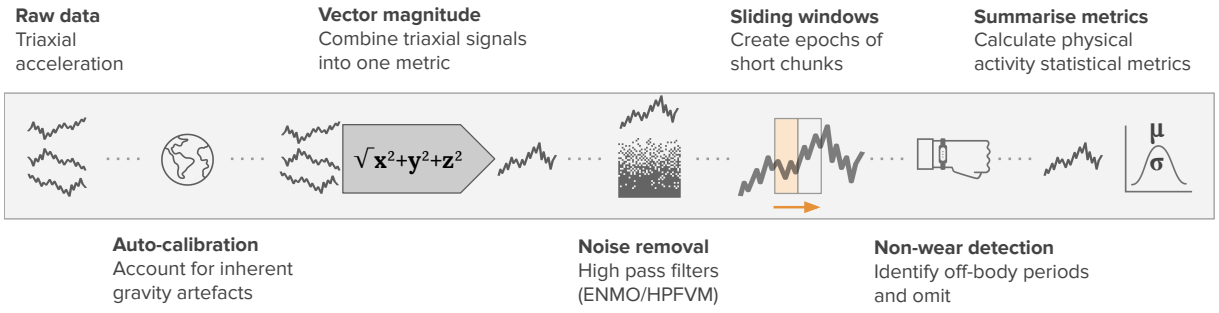


Figure 2.5: **Typical feature engineering pipeline for movement sensor data**. From raw accelerometer data to appropriate filters and summary statistics. (*ENMO*: Euclidean Norm Minus One, *HPFVM*: High-pass Filtered Vector Magnitude)

nomena like the seasonality or repeatability, but might require extensive computation (Fulcher, 2018; Fulcher and Jones, 2014).

The field of signal processing and waveform analysis has proposed multiple algorithms for extracting useful features from sensor timeseries. A common first step is computing a signal quality index (SQI) which assesses how physiologically reasonable a signal is (Orphanidou et al., 2014). An SQI is typically a combination of thresholds that should be met in order to keep or discard a signal (for example, in the case of PPG signals, check that all RR[6] intervals are < 3s). We can group the techniques for extraction of a waveform signal into two categories: filter based or feature based. Filter-based techniques are performed in a single step and they attenuate frequency components of the raw signal (e.g., bandpass filtering, wavelet transforms, detrending, or resampling). Feature-based techniques involve multiple steps and consist of extracting windowed feature measurements (e.g., amplitude, kurtosis between peaks, or maximum upslope). Another useful grouping is between time-domain and frequency-domain features: the former usually operate on the raw signal and calculate peaks or zero-crossings[7] over time, while the latter operate on the power spectrum of the signal using Fourier and wavelet transforms, or auto-correlations for periodicity estimation. We should note here that every application area has its own domain-specific features that attempt to best leverage the underlying data. We point the interested reader to indicative papers discussing the role of features for breathing rate estimation from PPG (Charlton et al., 2017) and human activity recognition with accelerometers (Haresamudram et al., 2019).

After the calculation of the appropriate features, they are fed to ML models. If static metadata exist (like demographics or personality traits), they are concatenated with the sensor features on a large feature vector. Common classification algorithms that appear in the literature are the Logistic Regression, Random Forests, Support

[6] the time elapsed between two successive R-waves of the QRS signal on the electrocardiogram
[7] a point where the sign of a signal changes (e.g. from positive to negative)

Vector Machines, and variants of Neural Networks. When the number of features increases as a result of the extensive feature extraction, we might end up with sub-optimal results: learning algorithms under-perform when the number of features is higher than the number of samples *("the curse of dimensionality")* (Friedman et al., 2001). As a result, we aim to reduce the number of features before training, either through feature selection or dimensionality reduction. For example, in a study of recognition of state changes in bipolar patients with smartphones (Grünerbl et al., 2015), data is reduced using Linear Discriminant Analysis. More robust approaches include the Principal Component Analysis (PCA), however we see cases such as this study on stress recognition (Bogomolov et al., 2014), which avoided using it because the transformation produces new variables that are difficult to interpret. The trade-off of interpretability and feature representation is an important open problem in this area, which is particularly exacerbated by the dominance of neural networks.

Most notably, and reiterating on the curse-of-dimensionality argument, the main bottleneck of using either more sophisticated models or features, is the risk of overfitting. Especially when dealing with human-generated data, overfitting refers to models which memorize exactly a particular set of data, but fail to generalize to a different population. As mentioned in previous sections, many studies conducted in the area of mobile health, included small numbers of participants. For example, some milestone studies using mobile sensors to track mood, sleep or other behavioral outcomes, analysed data from around 10 to 50 participants (Lane et al., 2011; LiKamWa et al., 2013; Wang et al., 2014). Based on each study's setup, the required granularity, and the duration of the tracking, each participant might contribute some tens or hundreds of data-windows. It goes without saying then that only the simplest models and features can be effective in this low data regime. In this thesis, we extensively discuss the trade-offs of feature engineering and more automated approaches for longitudinal sensor data, as well as limitations of encoding domain knowledge when processing raw data.

2.4.2 Raw sensor time-series modeling

The state of the mobile sensing–ubiquitous computing research community until recently used to resemble the computer vision community (then *image processing*) around 10 years ago. Due to the inability of algorithms back then to work directly on the raw pixels of an image (raw sensors in our case), researchers published inventive methods that were called *feature descriptors*. Seminal papers of that time like the *Scale Invariant Feature Transform* (SIFT) (Lowe, 2004), or *the Histogram of Oriented Gradients* (HOG) (Dalal and Triggs, 2005), are handcrafted algorithms that extract interest points

from an image based on geometry[8]. The turning point was in 2012 when the *Imagenet* study (Krizhevsky et al., 2012) showed that better results are possible with deep learning that does not need all these extra hand-crafted features.

The *Imagenet moment* in mobile sensing has yet to come because of various reasons. Datasets are not as big as to be fully exploited with deep learning, as well as there are no big benchmark datasets that are systematically evaluated through yearly competitions. Also, unlike object recognition, there is not *one* established task; we see many overlapping but different areas covering mood, stress, schizophrenia, bipolar disorders, sleep patterns, social interactions or depression, to name a few.

However, this plethora of tasks is achieved with a small number of sensors that reside in mobile phones or wearables (motion, ECG, etc), which can be modeled with light pre-processing in a similar manner with recurrent or convolutional neural networks. The field of human activity recognition embraced these methods early due to the availability of limited but standardized benchmark datasets and showed state-of-the-art results (Hammerla et al., 2016). This quickly spread to other tasks such as in mental-wellbeing modeling (Cao et al., 2017), and a more unified architecture was presented in *DeepSense* (Yao et al., 2017). This architecture integrated convolutional and recurrent layers and showed the efficacy of CNNs to learn local patterns, and RNNs to learn temporal properties. They demonstrated that the same architecture can generalize to various domains: car tracking with motion sensors, human activity recognition, and user identification with biometric motion analysis.

In the past three years, the literature has converged to the use of 1D CNNs (and sometimes a combination of CNNs and RNNs), when processing large datasets (Ballinger et al., 2018; Saeed et al., 2019). However, in tasks where non-standardized or smaller datasets are employed, researchers still use traditional features and linear models (Schubert et al., 2020). It is then obvious that we need better methods for "small data". As discussed earlier, the focus seems to have shifted to better training paradigms which exploit unlabeled data in a more efficient manner. In 2021, we would argue that the field of mobile and wearable health can skip the *Imagenet moment* altogether and strive for achieving its *BERT moment* (Devlin et al., 2019). Just as Imagenet showed that deep learning works in supervised setups, BERT showed that simple pre-training (fill the gaps) on massive cheaply-sourced unlabeled data produces models which can transfer effectively.

This thesis takes a pragmatic approach with regards to deep learning and feature engineering. In the next chapter, we see that sometimes careful feature engineering

[8]It is noteworthy – and somehow ironic –, that the vision community is now using similar transformations as targets in self-supervised pre-training tasks. We might have come full circle in that instead of inventing features, we now do so with tasks, although it highlights that the same operations capture very fundamental information.

achieves results on par with deep learning, especially when it comes to modeling very long time-frames of fine-grained signals. Last, we argue that, beyond accuracy, deep learning is particularly useful for its internal representations and the re-usability thereof in other transfer tasks.

2.5 Relating to planned work

Having reviewed the state of the art in the area of ML for mobile health with a focus on mental and physical well-being, we will now discuss how the next chapters improve upon existing works. We acknowledge that the building blocks of machine learning such as neural network layers (CNNs or RNNs for example) have become commoditised and can be used in various configurations to create ”novel” architectures. However, our focus has been on targeted approaches which best leverage large retrospective behavioral data, such as new loss functions that are inspired by the data generation distribution (see Chapter 5), or ways to visualize neural activations towards understanding the temporal dynamics of the underlying data (see Chapter 4).

Regarding mental health, we present results on a global dataset of mood reports associated with passive sensing data, orders of magnitude larger than other studies (LiKamWa et al., 2013; Lane et al., 2011; Taylor et al., 2017). Also, given the impracticability of obtaining clinical labels — such as depression (Suhara et al., 2017; Shah et al., 2021) — in this scale, we focused on mood prediction whose instability is a predictor of poor mental health. Chapter 3 proposed a methodology which combines clustering historical mood trajectories with classification models that operate on these clusters. A conceptually similar methodology follows a cluster-then-classify approach on a smaller sample (Taylor et al., 2017), although their goal is to provide personalized predictions to these clusters. Our work provides insights in terms of evaluating the impact of adding sensor data to personality metadata through ablation studies, as well as providing actionable recommendations on handling sparse smartphone sensor data. Chapter 4 proposed an encoder-decoder multi-task model for mood forecasting, which to the best of our knowledge has not been investigated before. Mood forecasting was the subject of other studies (Suhara et al., 2017), however they focused on using multiple mobile sensor data for this task. Our aim is to understand the interplay of the two dimensions of mood through multi-task learning and investigate the learned temporal dynamics of mood as well as how they relate to personality or external factors.

For the second half of the thesis which is related to physical health, we present results on a retrospective dataset of high-frequency free-living physical activity sensor data and lab-measured clinical outcomes. Chapter 5 proposed a model that — by

mapping activity to future heart rate — learns a meaningful representation that can be used as features in many downstream tasks. Many studies have focused on unimodal unsupervised learning on either activity or ECG data (Saeed et al., 2019; Tang et al., 2021; Sarkar and Etemad, 2019), however our work was one of the first ones proposing a multi-modal pre-training task. Our work is conceptually similar to (Hallgrímsson et al., 2018), with the main difference of this approach being the requirement of a historical input of one month of data, as well as the application to more coarse-grained data than ours. On the other hand, our approach provides ablation tests and experiments with a wider range of transfer learning outcomes. It is also noteworthy that ideas from our model are incorporated in a recently published paper proposing a self-supervised model for forecasting adverse surgical events (Chen et al., 2021). In Chapter 6, we develop models which can predict the gold standard of cardio-respiratory fitness using completely free-living wearable data along with demographics. Once again, the difference to other studies is in the scale and validation methods. While recently published studies used 50 (Eades et al., 2021), 46 (Altini et al., 2016), 37 (Bonomi et al., 2020), and 191 patients (Kwon et al., 2019), we leverage a cohort of over 11, 000 participants with a large sample of over 2, 000 people repeating the protocol almost a decade later. This allows us to train robust models which can potentially generalise to the population level. Also, we are able to investigate hypotheses pertaining to fundamental challenges in machine learning such as distribution shifts, by applying the same models to the longitudinal cohort.

Last, we should note that the publications arising from the chapters of this thesis have already been cited by other researchers more than 30 times, with multiple studies building on top of our work.

Chapter 3

Multimodal mobile sensing for mood prediction

Το όλον είναι μεγαλύτερο από το άθροισμα των μερών του[1]
–Aristotle

3.1 Introduction

In this chapter, we present a training pipeline for population-scale mobile sensor data towards more accurate mood clustering and prediction. These results motivate the complimentary use of different modalities through multimodal learning, which is further studied in Chapters 5 and 6.

Experience sampling has long been the established method to sample people's mood in order to assess their mental state. Smartphones start to be used as experience sampling tools for mental health state as they usually accompany individuals throughout their day and can therefore gather in-the-moment data. However, the granularity of the data needs to be balanced with the level of user inconvenience that these tools introduce. Interrupting users during their daily lives at a high frequency and with the same purpose is seen as a high burden by many users (Mehrotra et al., 2015), as it is evidenced by the high dropout rates reported in these applications. Indeed, according to recent statistics, more than two thirds of people who download a mobile health app used it only once (Lee et al., 2018). As a consequence, the data collected with this technique is often sparse. This has been obviated by the use of passive sensing in addition to mood reports; however, this adds additional noise. In this chapter, we show that psychological traits collected through one-off questionnaires combined with passively collected sensing data (movement from the accelerometer

[1]The whole is greater than the sum of its parts.

and noise levels from the microphone) can be used to detect individuals whose general mood deviates from the common *relaxed* characteristic of the general population. By using the reported mood as a classification target, we show how to design models that depend only on passive sensors and one-off questionnaires, which do not entail the user inconvenience associated with experience sampling. We validate our approach by using a large dataset of mood reports and passive sensing data collected in the wild with tens of thousands of participants, finding that the combination of these modalities achieves the best classification performance, and that passive sensing yields a +5% boost in accuracy. We also show that sensor data collected for a week performs better than single days for this task. We discuss feature extraction techniques and appropriate classifiers for this kind of multimodal data, as well as the overfitting shortcomings of using deep learning to handle static and dynamic features. These findings have implications for mobile health applications, that can benefit from the correct modeling of passive sensing along with additional user metadata.

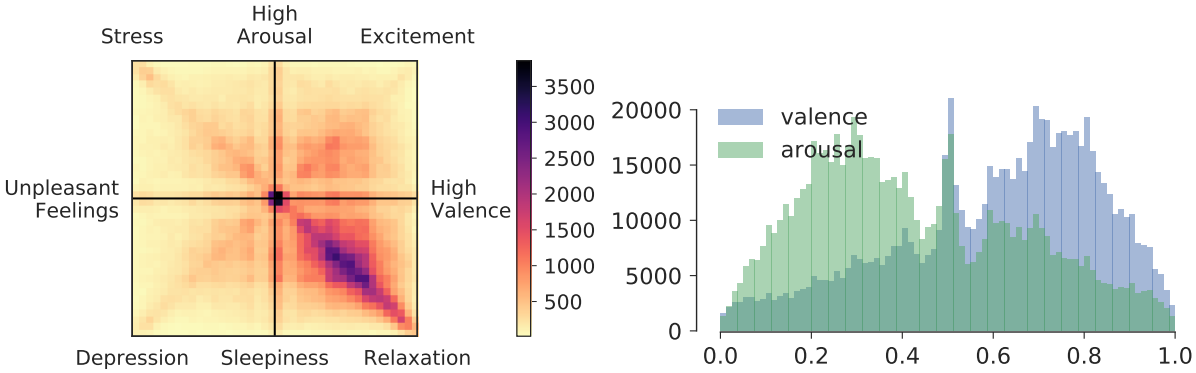
A summary of the contributions of this chapter is as follows:

- We conducted an extensive data exploration of the self-reported moods provided by 17,251 of the users of an experience-sampling based smartphone application, with the aim of identifying the most common reporting behavior in order to characterize *mentally healthy* individuals in the context of our research. Our findings showed that the majority of the population in our dataset reported –on average– a relaxed feeling (bottom-right side of the affect grid, Fig. 3.1a), which is in line with previous research (Russell et al., 1989).

- We provide a supervised learning methodology to detect individuals whose general mood deviates from the common *relaxed* mood distinctive among mentally healthy individuals (Russell et al., 1989). Our methodology does not involve any kind of cumbersome experience sampling, but only uses one-off questionnaires (demographics, personality, etc.) as well as sparse and noisy passive sensing data collected with the accelerometer and microphone sensors of individuals' smartphones.

- We performed an extensive evaluation of our methodology using a large scale dataset collected in the wild. Our results showed that the combination of one-off questionnaires and passive sensing data gives the best performance in mood prediction. Indeed, by adding passive sensing data we achieved a +5% in accuracy (75% in absolute) with respect to only using questionnaires.

These findings have the potential to inform future developers of mobile health applications as well as psychologists on how to how to most effectively use one-off



(a) Heatmap of valence and arousal.

(b) Distribution of valence and arousal.

Figure 3.1: **Aggregate 735.778 self-reported mood scores in the EmotionSense dataset collected from 17.251 users**. Most users report neutral (around 0.5,0.5) and calm-happy (bottom-right quadrant) mood on the affect grid (a). The two multimodal distributions (Pearson's r=-0.23, $p < 0.001$) of the mood (b).

questionnaires and passive sensing data for the early detection of symptoms of mental disorders at scale.

3.2 Problem formulation

Hypothesis

A machine learning model that combines passive sensor signals with traditional user-level metadata should be more accurate in predicting mental health outcomes (mood differences).

Mobile health applications, aimed at assisting users with their mental health to prevent clinical intervention outcomes should minimize the burden to the user so as to increase adherence and satisfaction with the app. Instead of the timely and continuous collection of mood self-reports, psychological traits obtained through one-off questionnaires, as well as passive sensing data, should be preferred in order to design effective and useful applications. Our aim in the rest of this chapter is to investigate how psychological traits and passive sensing data can be used to detect individuals who might not feel mentally well, i.e., users who have been reported moods that deviate from the general reports of the population.

To do so, we first conduct an exploratory analysis of the mood reports provided by more than 17,000 individuals for a period of more than 3 years, in order to identify the most common set of mental states (moods) reported by any of these individuals (Section 3.3.1). Given the scale and the real-world context of the data collection, we believe our results are general enough to be representative of the whole population.

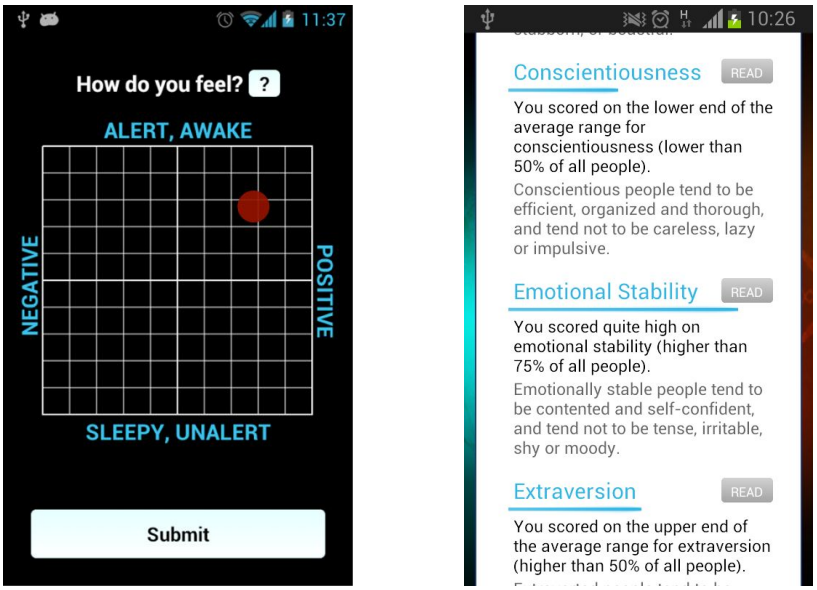


Figure 3.2: **Screens of the mood tracking application**. Users could report their mood in an affect grid, as well as complete personality-related and other questionnaires.

We then use these findings as the ground truth to validate our machine learning methodology and identify individuals whose record of reported moods deviates from that of the majority, by using only one-off questionnaires and passive sensing data (Section 3.3.2). We provide further details of the data used in our analysis and experiments in the rest of this section.

3.2.1 Data modalities

Experience sampling. Figure 3.1 shows the aggregate of mood self-reports for all users of the application, where the bottom-right quadrant, corresponding to *relaxed* mood, is the most densely populated, a result that matches previous studies in the area (Russell et al., 1989). Due to the real-world context of the data collection, users did not always report their mood even if they were prompted to do so, which might be consequence of the burden that experience sampling brings to the users. Figure 3.3 shows in more detail the CCDF of moods reported per participant, included the reports they were *expected* to do given the time they were using the app, the reports they were prompted to do but did not give *(missed)*, and those that they actually *did* give. The results show that alternatives to experience sampling are required to design effective, long-term, mobile health applications for mental health. As we will show later, by using the reported mood as a classification target we can design systems that depend only on passive sensors and one-off surveys.

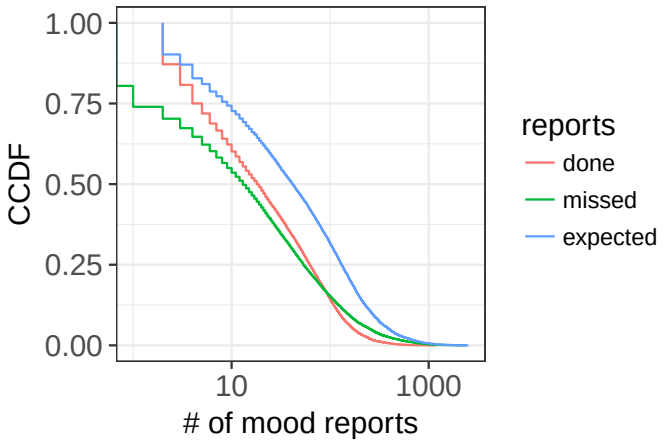


Figure 3.3: **Complementary cumulative distribution function of the mood reported by users during the time they were using the application**. This includes (i) the self-reports actually done (*done*), (ii) those that users were prompted to report but they did not do so (*missed*) and (iii) the sum of both (*expected*).

One-off questionnaires. Previous research has found a link between self reported mood and personality traits such as emotional stability (Ching et al., 2014; Geukes et al., 2017). However, to the best of our knowledge, it is not clear yet how to utilise personality, and other psychological traits, to detect potentially mentally unhealthy individuals. In the *EmotionSense* dataset, a subset of the users (12, 106, 70% of total) completed some one-off surveys providing information regarding their demographics, personality, gratitude, health, sociability, job satisfaction, life aspirations, connectedness, and satisfaction with life.

Passive sensing data. Data collected through the built-in accelerometer sensor of our smartphones provide valuable insights into our activity level throughout the day. At the same time, previous research has demonstrated the link between activity level and happiness (Lathia et al., 2017; Servia-Rodríguez et al., 2017). We hypothesize that our activity level throughout the day has a high impact on how we feel on that day and therefore we use this data in our experiments. In the *EmotionSense* dataset, accelerometer samples consist of $[x, y, z](m/s^2)$ axes data for periods of 5, 8 or 10 seconds, collected at different intervals throughout the day depending on the version of the application. Microphone samples, on the other hand, provide insights into the noise level in the user's environment. As with activity, we hypothesize that how we feel (our mood) influences/is influenced by the kind of places or environments we visit and the level of noise in these spaces. Therefore we use this in our experiments. To preserve privacy, the *EmotionSense* application only recorded the amplitude level of noise at $20Hz$ (the lower limit of human's audible spectrum) for periods of 5, 8, and 10 seconds at different intervals throughout the day depending on the version of the

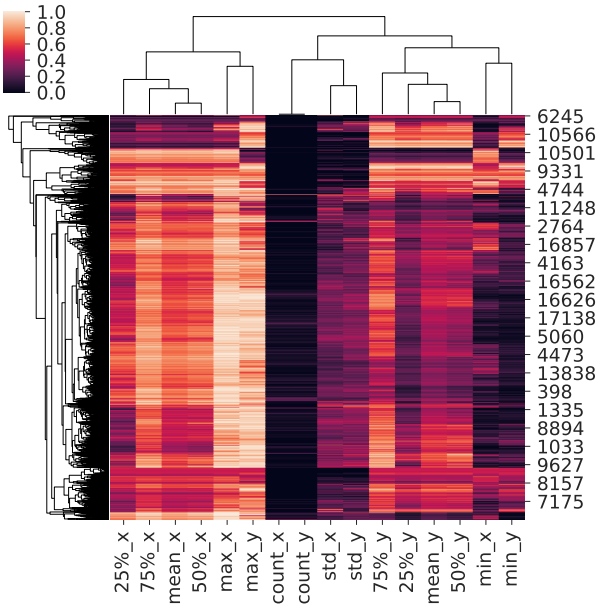


Figure 3.4: Hierarchical clustering of the users (y-axis, only some user IDs are visible) and features (x-axis) extracted from their historical mood (_x=valence, _y=arousal). The colorbar represents the actual value of the feature.

application.

Varying amounts of data is available for each of the sensors and self reports, mainly due to the uncontrolled methodology for participant recruitment. Also, the in the wild nature of the data collection makes the available data noisy and sparse, which adds to the challenge. We present more detail on how we dealt with the noise and sparseness, as well as on the number of participants and days of sensed and self-reported data used for each analysis, in Sections 3.3.1 and 3.3.2.

3.3 Method

3.3.1 Clustering historical trajectories

The main goal of our research is to investigate whether psychological traits and passive sensing data can be used to identify users whose set of mood reports deviates from those of the general population, which might be indicative of some mental condition. Fig 3.1 shows a visualization of the aggregation of self-reports provided by users in the *EmotionSense* dataset, where the most common mood reported is in the bottom-right side of the affect grid, corresponding to the *relaxed* mental state. However, it is not clear how to split the affect grid into bins or classes. We propose not to hard code any thresholds and potentially induce biases in our labels, but instead to rely on clustering techniques in order to make labels naturally emerge from the data. The rest of this section describes in detail the methodology to label users into *relaxed*/non-*relaxed* in

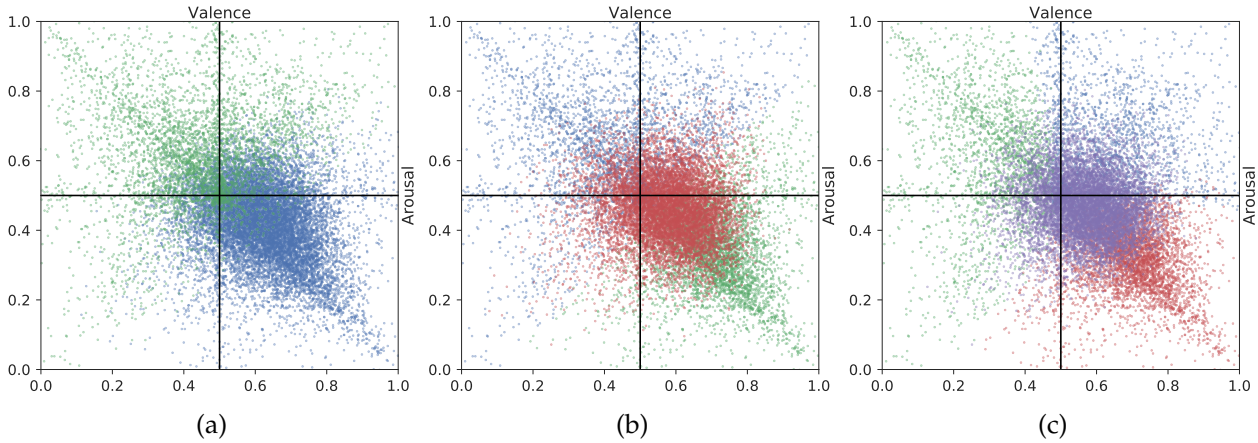


Figure 3.5: **Clustering the historical mood trajectories**. Grouping in 2, 3, and 4 clusters: (a,b,c) Affect grid plot of the mean valence and arousal of the clustered users (every dot is a user). The clusters of the first plot (a) are used as prediction labels for the subsequent classification task.

the *EmotionSense* dataset.

Feature extraction. A mood self-report in the affect grid is described by means of two coordinates: the x-coordinate that indicates the feeling in terms of its positive and negative affect, and the y-coordinate indicates the intensity of alertness. The history of mood-reports of an individual consists of time-series trajectories of [x,y] tuples recorded over time in the affect grid. Also, the in-the-wild setup is reflected on that (i) the number of self-reports reported by different individuals might be different, and (ii) that for a given individual, the reported moods might not be consecutive (as a consequence of users missing reports). In order to cope with this variability and obtain independent features to allow clustering algorithms to learn representative clusters, we extract eight simple features for each axis or coordinate, namely *counts*, *mean*, *std*, *min*, *max* and *quantiles* (25%, 50%, 75%), resulting in 16 final features for every user. Missing values are replaced with zeros and *minmax* [0,1] normalization is applied to the final features column-wise. Due to the sparsity of the mood and the power law distribution of the counts, these two *count* features that measure non-missed reports are affected the most by the normalization, concentrating all their mass close to zero.

Clustering. We then apply a clustering algorithm to produce mutually exclusive clusters (*k-means* (MacQueen et al., 1967)). In order to come up with the optimal number of clusters, we apply the Elbow method (Thorndike, 1953) where we increase the number of clusters and observe the drop of the evaluation metric. Here, we use the silhouette metric (Rousseeuw, 1987) which measures how similar a sample is to its own cluster in comparison with other clusters. Other clustering algorithms might also be used. In fact, techniques such as *hierarchical (agglomerative) clustering* (Kundaje

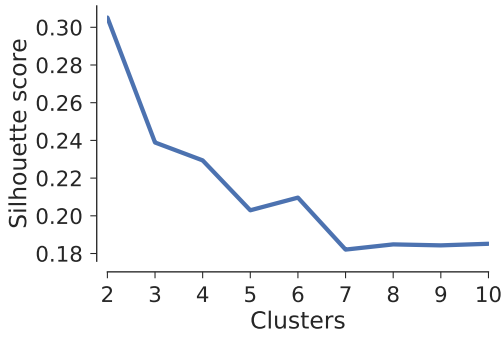


Figure 3.6: **Clustering evaluation**. *Elbow* plot to determine the optimal number of clusters, estimated with the silhouette score.

et al., 2015) applied to the matrix of [users,features], can be used to find partitions in the data, but also to uncover overlapping patterns between features.

We applied our methodology to identify non-*relaxed* users (or those that deviate from the most common mood feeling reported) in the *EmotionSense* dataset. For each of the 17,251 users who have reported their mood at least once, we obtain 2,682 sparse mood reports completed over 3 years, for valence and arousal. This is the final sample we used for this experiment.

Exploratory analysis. As a first exploratory analysis, we apply hierarchical clustering to the historical mood of the users. Figure 3.4 shows the resulting trees. We observe that there are multiple user groups shown on the left side tree, pointing out that some mood reporting behaviors resemble those of other users. However, it is not easy to spot a clear relationship due to the number of users. The highest-level clusters predictably split to valence and arousal features. However, there are some *inliers* in those clusters: for example, the maximum arousal (max_y) belongs to the valence cluster while the counts (counts_x) and the minimum (min_x) of valence goes into the arousal group. These feature clusters provide hints regarding the non-linear relationships of the mood components.

Label extraction. After applying k-means we obtain the labels to use in our experiments. We repeat the experiments by varying k, that is the resulting number of clusters. Figure 3.5 shows the resulting clusters when increasing the number of clusters from 2 to 4. For 2 clusters (Fig. 3.5a), by plotting the mean valence and arousal in the affect grid, we notice a group of consistently relaxed users on the bottom-right quadrant and another group that consists of depressed, stressed and excited users on the rest of the grid. When we further increase the number of clusters, the classes are not so apparent. For example, with 3 clusters (Fig. 3.5b) we spot a central neutral group which is now distinct, while the rest is similar to the previous plot (relaxed and non-relaxed). Finally, for 4 clusters (Fig. 3.5c), we spot again the middle neutral

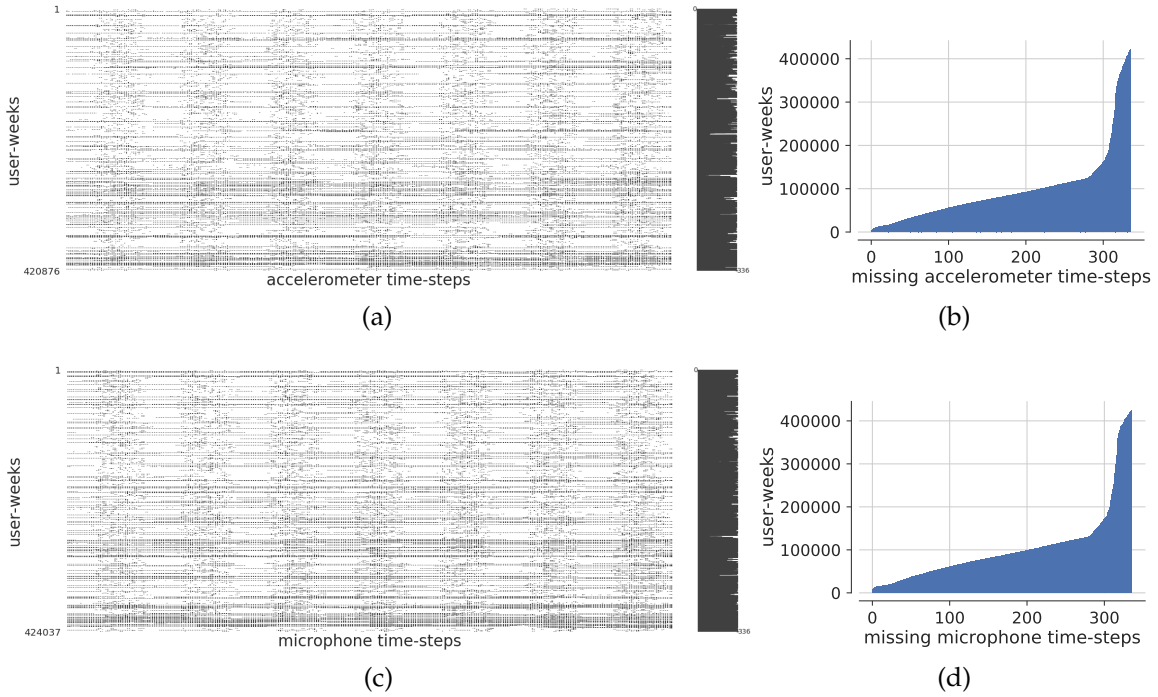


Figure 3.7: **Sparsity analysis of the sensors**. Missing values for the sensors on the weekly (a,c) level. Cumulative distribution functions (CDF) for the missing time-steps (b,d) show the long tail distribution of sparsity. Some daily periodicity is also spotted. Similar conclusions are drawn with the daily level sensors.

users but this time the valence axis breaks down to two areas: excitement (up right) and relaxation (down right). It notable that the negative feelings (left side) do not break down to sub-clusters, hinting that the two dimensions of arousal for unpleasant feelings (stress and depression) might share some common characteristics. However, these plots (3 and 4 clusters) present significant cluster overlap.

Finally, we perform the elbow method to quantitatively find the optimal number of clusters. Figure 3.6 shows that the top silhouette score is 0.30 (the higher the better) with two clusters, while it drops to 0.23 with three clusters. We observe that it plateaus at around 0.20 with seven clusters or more. These two groups will be used as a label in the machine learning pipeline to infer non-*relaxed* users from one-off questionnaires and passive sensing data in the next section. We are aware that these clusters are inferred information and thus could include some errors, however we incorporate the silhouette score with the lowest error. Please note that there is a class imbalance between the clusters on the *user* level: cluster 1 (65%), cluster 2 (45%), which we will address later in the section.

3.3.2 Classifying clustered mood

We now describe our methodology to identify non-*relaxed* individuals from their psychological traits obtained through one-off questionnaires, and passive sensing data collected using the accelerometer and microphone sensors of their smartphones. We follow the workflow in Figure 3.8, where we begin by extracting features from the accelerometer and microphone raw data, as well as one-hot encoding the answers of user-surveys. We then perform a two-step feature selection, where we first calculate the feature significance of a real-valued feature to a binary target as a p-value using the univariate Mann-Whitney U test (Mann and Whitney, 1947), and then transform these selected features with Principal Component Analysis (PCA) (Pearson, 1901) to obtain feature combinations with the maximum variance. These features are finally fed to classifiers. We detail these steps below.

Questionnaires. One-off surveys cover a wide range of a user profile attributes such as demographics, personality, gratitude, health, sociability, job satisfaction, life aspirations, connectedness, and satisfaction with life. These 92 features are represented as Likert-scales or categories which are described in detail in (Lathia et al., 2017). In order to be appropriate for machine learning models, the categorical features are transformed to individual features with *one-hot encoding*, so that a feature with e.g. 3 possible choices (Yes, No, missing), is transformed to 3 different features. Categorical features include the gender, age group, education level and ethnic group among others. The total list of questionnaire features is 131.

Accelerometer. We consider the 3 (x,y,z) dimensions of the accelerometer and compute the magnitude of the acceleration for 5, 8, and 10-second samples, resulting in 48 time-steps for every user-day (336 time-steps for every user-week). We aggregate the sensor in 30-min bins since this level of granularity is the best trade-off between data sparsity and modeling the sub-hourly movement of individuals. By doing this light processing, we end up with one time-series instead of three, combining the three axes into one time-series. Based on the sparsity histogram (Fig. 3.7b), we filter those samples that have at least 50 time-steps during the week (20 time-steps during the day). This time-series is normalized with *minmax scaling* to a [0.05-1] range and the missing values are replaced with zeros. We extract 721 simple and second order features that cover a wide range of attributes of a sensor such as the energy, auto-correlation, entropy, trends, wavelet and Fourier coefficients, peaks, etc. For a comprehensive list of the features we refer the reader to the documentation of the *tsfresh* library (Christ et al., 2018) and the Appendix A.

Microphone. Similar to the accelerometer data, we compute the mean of the 5, 8, and 10-second window over the initial raw microphone data over the amplitude level of noise at 20 Hz, ending up with 48 time-steps for every user-day (336 time-steps for

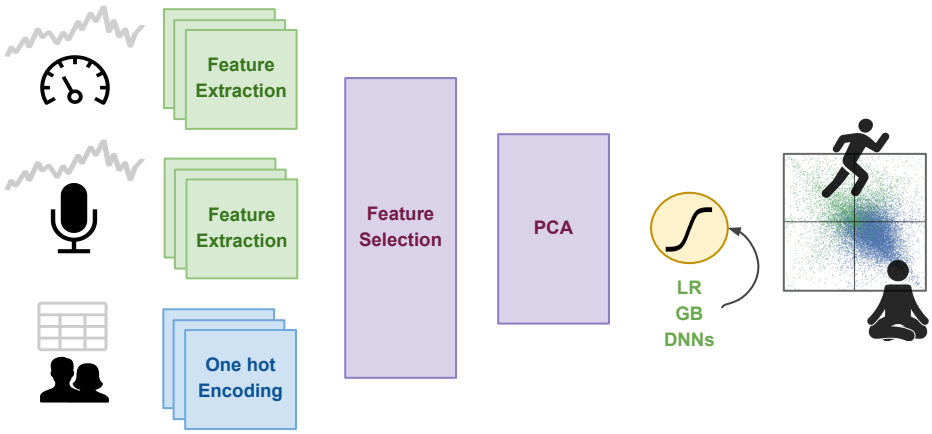


Figure 3.8: **Data flow**. Conceptual illustration of the data modality merge and pre-processing steps for the mood prediction task.

every user-week). We apply the same filtering, normalization and feature extraction as the accelerometer above, resulting in 717 features.

Seasonality. Temporal features are extracted by the end of the sensor user-week timestamp in order to capture the inherent seasonality patterns. Namely, we compute these 5 increasingly detailed time-aware features: the number of the calendar year quarter, month, week, day of week and hour of day. We group these features under the *sensor* modality that we introduce later.

Classifiers. We considered three different classifiers for our inference task: Logistic Regression, Gradient Boosting Trees and a Deep Neural Network. Below we describe the details of our implementation.

Logistic Regression (LR). An *sklearn* implementation of a binary logistic regression, with penalty of $L2$ regularization along with a $C = 1$ (inverse of regularization strength), was tested.

Gradient Boosting Trees (GB). An *sklearn* implementation of a gradient boosting was tested. Reportedly the state-of-art in feature-based machine learning (Olson et al., 2018), this classifier forms an ensemble of weak prediction models, typically decision trees.

Deep Neural Network (NN). We use a straightforward bottleneck architecture of 4 feed forward *Dense* layers of dimensionality 100-50-100. The reduced dimensionality in the middle (50 units) has been shown to lead to better generalization in deep learning architectures (Lozano-Diez et al., 2017; He et al., 2016). A rectified linear unit (*ReLU*) (Glorot et al., 2011) activation is applied at the output of every layer, followed by a *batch normalization* layer that transforms the output to have zero mean and unit variance (Ioffe and Szegedy, 2015). *Dropout* of 50% probability is applied to every layer to reduce overfitting (Srivastava et al., 2014). The final layer performs a *softmax* activation which estimates the cross-entropy loss, while the backpropagation optimizer is *Adam* (Kingma

Table 3.1: **Classification performance**. AUCs for the task of predicting mood group based on weekly or daily sensors, across 10 cross-validation runs along with standard deviation in brackets (NN=neural network, LR=Logistic Regression, GB=Gradient Boosting).

Modality	Weekly			Daily		
	LR	GB	NN	LR	GB	NN
Sensors (S)	0.575 (0.03)	0.555 (0.03)	0.550 (0.04)	0.543 (0.04)	0.514 (0.02)	0.510 (0.03)
Questionnaires (Q)	0.690 (0.05)	0.627 (0.10)	0.687 (0.09)	0.671 (0.11)	0.729 (0.09)	0.701 (0.09)
All ($S + Q$)	**0.749** (0.06)	0.721 (0.03)	0.725 (0.06)	0.706 (0.07)	**0.740** (0.09)	0.697 (0.10)

and Ba, 2014). We train for 300 epochs or until the validation loss stops improving for 10 consecutive epochs. Our implementation is based on Tensorflow/Keras. In Appendix A we provide more details regarding the models.

3.4 Evaluation

We now detail the evaluation of our methodology to identify non-*relaxed* users from one-off questionnaires and passive sensing data described in Section 3.3.2. We used the *EmotionSense* dataset, for our experiments, and the clustered mood we obtained using k-means in the Section 3.3.1 as the labels for the classifiers. Below we indicate how we merged the data from the different modalities and how we partition the dataset for our experiments.

3.4.1 Experimental setup

Modality merge. Experiments in the wild such as this one do not guarantee complete and fine-grained data, especially when they involve battery consuming tasks such as sensor-tracking or input-based prompts such as self-reports from users. Therefore, not all modalities appear for the same users. We start by merging the accelerometer and microphone modalities, resulting in 141,261 user-weeks while we concatenate their features along with the seasonality features. Next, we find which users from those weeks have completed at least a single questionnaire and concatenate these *static* features to the feature vector, resulting in 131,793 user-weeks. Finally, we merge with the clusters that we produced in the previous section, so that every user-week feature vector corresponds to one of the two user mood clusters. Please note that these clusters came up by taking into account the full mood history of the users and therefore we do not imply that mood is static. Predictably, the high class-imbalance on the *user* level seen earlier is exaggerated here because only 7% of the *user-weeks* belong to cluster 2 (green in Fig. 3.5). As a result, we subsample the majority class, resulting in 18,998 balanced user-weeks from 2,812 users. The same processing is followed for the daily

sensors: 112,161 user-days after sensor merge, 106,672 after questionnaire merge, and we end up with 16,470 user-days from 1,859 unique users when we merge with the labels and sub-sample.

Feature ablation studies. In order to identify which feature modality contributes more to the classification we repeat our experiments with 3 different modalities: only sensors (accelerometer, microphone and seasonality), only one-off questionnaires (psychological profile), and a combination of these. To make for a fair comparison, for every modality we keep only 100 features that we feed to the classifiers. Since every modality contains different numbers of features (combined=1,564, sensors=1,434, questionnaires=130), we perform a two-step feature selection. First, we calculate the feature importance with a Mann-Whitney U test (Mann and Whitney, 1947). Next, these selected features are transformed with Principal Component Analysis (PCA) (Pearson, 1901), a common decorrelation method that produces latent features, resulting in 100 components.

User based cross validation. Typical cross-validation would not be adequate in our task since some *static* features such as the age or gender are repeated for different weeks because they belong to the same user. Therefore, we create training and test sets from *disjoint* user splits, making sure that weeks from the same user do not appear in both splits. Please note that this does not result in perfectly balanced class splits, but the evaluation metric we are using, the Receiver operating characteristic-Area Under Curve (*ROC-AUC* or simply *AUC*) is robust to class imbalances. Even using this metric, it is not easy to guarantee that a split picked a representative test-set, thus we perform a 10-fold-*like* cross validation using 10 different seeds to pick disjoint users. Consequently, we conduct an extensive experimentation by testing 180 models (3 modalities $\times$ 10 user splits $\times$ 3 classifiers $\times$ 2 temporal levels). The size of the test set is 10% of the dataset, and of the rest 90% used for training we keep a random 10% for validation (used only in neural networks). This validation set belongs to the same distribution as the training set. We report the average performance of the folds and the standard deviation.

3.5 Results

We now present the classification results of predicting whether a user-week/day belongs to the relaxed or to the rest of the mood spectrum, based on sensors, questionnaires and other meta-data. As discussed earlier, we performed extensive experiments and trained 180 models to evaluate the impact of the different modalities and user splits. In Table 3.1, we present the mean classification performance of the experiment setup described in the previous section, that of predicting the mood cluster group (relaxed

or not) based on each user's weekly/daily sensors and questionnaire metadata.

Week level. By using the sensors at the week level we achieve the best overall performance of 0.749 AUC, which comes from the LR model, while the NN comes second with 0.725. Even though the NN and GB are non-linear classifiers, they underperform, possibly due to the issue of overfitting or the data compression with PCA. Also the LR model shows stability with the lowest standard deviation across all cross-validation runs. Regarding the modalities, in the best case of the LR, the combined representation of the sensors and the questionnaires outperforms the single modality of questionnaires by +5.9% AUC and reaches +9.4% in the case of GB (with a lower max AUC in the combined representation though). The sole use of sensors achieves less than 60% for all the models. This ranking is consistent for all the classifiers.

Day level. Considering only one day of sensing data, the absolute results are slightly lower than those of the weekly level. Here, the GB model achieves an AUC of 0.740, while the LR comes second with 0.706. The NN presents similar performance for the combined and questionnaire representation, hinting that the daily sensors do not contribute significantly to it. However, the rest models show a rise of +1.1% (GB) and +3.5% (LR) in AUC, when we add the sensors to the questionnaires.

3.6 Discussion

These results show that by adding passive sensing to traditional personality and demographics surveys we are able to predict the mood group of individual users with a higher precision. Specifically, for our task we achieve $\sim$ 75% AUC by classifying users' status into relaxed or not. Also, we observe that by tracking the users for a longer time duration (one week, instead of one day), we achieve better performance. In hindsight, this is intuitive, since movement and noise levels are expected to be related with relaxation levels. Beyond the binary task, additional experiments with 3 or 4 clusters (multi-class) yielded worse results due to the significant cluster overlap and fewer data-points per class to learn. Furthermore, putting our results in the context of related work we see that similar datasets yield lower accuracy (around 65%) for slightly different tasks such as predicting tomorrow's mood (Taylor et al., 2017) or daily mood average (LiKamWa et al., 2013). Furthermore, perhaps the most closely related work to ours is the *Snapshot* (Sano, 2016) study. This study investigated how daily behavior gathered through passive sensing data influence sleep, stress, mood, and other wellbeing-related factors. Multiple papers focused on different aspects of the collected dataset, such as personalization with multi-task learning to predict tomorrow's mood, stress, and health (Taylor et al., 2017), prediction of happy/sad mood based on sleep history (Sano et al., 2015), or a denoising autoencoder to fill

in missing sensor data for mood prediction (Jaques et al., 2017). Similarly to our methodology, they first cluster the users before going into classification (Taylor et al., 2017), although their goal here is to provide personalized predictions to these clusters. However, our models do not distinguish between healthy and depressed patients, but instead predict the clustered mood group which roughly correspond to relaxed or not-relaxed users. From a more practical perspective, personalized models are difficult to be deployed in a real world scenario, since they require training N personalized models, with N being the number of users. Even though previous research has shown that better performance can be achieved by averaging the individual model accuracies (Canzian and Musolesi, 2015; LiKamWa et al., 2013), no results are reported on unseen disjoint users. Here, we propose a robust ML pipeline and report results on a disjoint user set.

3.7 Conclusion

In this chapter, we showed how the pervasiveness of smartphones has converted them into experience sampling tools to collect people's mood in order to assess their mental state. However the granularity of the data needs to be balanced with the level of user inconvenience these tools introduce on users' activities, which often results into very sparse data. In this chapter, we propose a machine learning methodology to detect if an individual's perceived mood differs from that of the general population, by solely considering their psychological traits collected through one-off questionnaires and passively collected mobile sensing data, thus avoiding the use of experience sampling questionnaires.

We evaluate our methodology by using a large-scale dataset collected in the wild for more than 3 years and 17,000 participants. An exploratory analysis of the data revealed that *relaxed* is the most common state reported by our population. Our experiments also confirmed that our methodology is able to distinguish between generally *relaxed*/non-*relaxed* individuals with a 75% AUC when using a combination of weekly sensors (accelerometer and microphone) and one-off questionnaire data (personality, demographics, etc) as inputs. Besides, the use of passive sensing data yields a +5% boost in accuracy. In a healthcare context, this accuracy suggests that we can group users correctly 3 out of 4 times using only short-time mobile phone sensing and sparse surveys. While this level of accuracy might not be adequate for deployment in clinical settings, our focus is mostly on the positive contribution of passive sensing.

This first empirical chapter of the thesis sets the tone of the following ones by introducing models and methodologies which can leverage mobile sensor data along with other traditional features. Although the employed models are commonly used in

the literature, here we propose a new robust model pipeline which includes clustering user trajectories followed by classification models. We consider a key contribution to be the correct handling of user-level data towards generalising to held-out populations. Last, regarding the application area, this chapter focused on the association of passive sensing to mental health, but we used aggregated clustered trajectories as mood outcomes. Therefore, we have not yet explored how mood variability across consecutive days could inform potential predictions of mental health issues. To this end, the following chapter puts forward a model which learns patterns in mood sequences through multi-task learning.

Chapter 4

Sequence multi-task learning for mood forecasting

Prediction is very difficult, especially if it's about the future.
–Niels Bohr

4.1 Introduction

In the previous chapter, we focused on the feasibility of large-scale mood prediction through passive sensing and proposed a multimodal training pipeline. In this chapter, motivated by the temporal dynamics of mood instability, we present an encoder-decoder model which exploits the bi-modality of mood with multi-task learning, enabling more accurate multi-step mood forecasting. These results motivate the use of forecasting as a means of learning meaningful representations, which is further explored in Chapter 5.

Smartphones are increasingly used as self reporting tools for mental health state because they accompany individuals throughout the day and can therefore gather temporally fine-grained data. However, the analysis of self reported mood data offers challenges related to non-homogeneity of mood assessment among individuals due to the complexity of individual mood states and the reporting scales that capture these, as well as the noise and sparseness of the reports when collected in the wild. In this chapter, we propose a new end-to-end ML model inspired by video frame prediction and machine translation, which forecasts future sequences of mood from previous self-reported moods collected in the real world using mobile devices. In contrast to traditional time series forecasting algorithms, our multi-task encoder-decoder recurrent neural network learns patterns from different users, thus allowing and improving the prediction for users with limited number of self-reports. Unlike traditional

feature-based machine learning algorithms, the encoder-decoder architecture enables to forecast a sequence of future moods rather than one single step. Meanwhile, multi-task learning exploits some unique characteristics of the data (for example that mood is bi-dimensional), achieving better results than when training single-task networks or other classifiers.

Our experiments using a real-world dataset of 33,000 user-weeks revealed that (i) 3 weeks of sparsely reported mood is the optimal number to accurately forecast mood, (ii) multi-task learning models both dimensions of mood — valence and arousal — with higher accuracy than separate or traditional ML models, and (iii) mood variability, personality traits and day of the week play a key role in the performance of our model.

This chapter makes the following contributions:

- We propose and adapt an end-to-end, stand-alone model inspired by video frame prediction (Srivastava et al., 2015) and machine translation (Sutskever et al., 2014), to forecast sequences of future moods — valence and arousal — from previous self-reported moods.
- Our evaluation on real world data reveals that (i) our model forecasts tomorrow's mood with ± 0.14 minimum error, and 7 days later with ± 0.16 error on the affect grid, and (ii) that the multi-task model trained to learn predicting both valence and arousal simultaneously is more accurate than independent models trained on each dimension separately, especially for arousal.
- We show the internal learned *black-box* representations of the deep neural networks and observe that different neurons learn different non-linear sequential patterns, which helps us to understand the complex trajectories of future mood.
- An exploratory post-hoc analysis reveals that the accuracy of the learned model is related to the day of the week, personality traits and mood variability. Specifically, our model performs better for *open*-personality users and on weekends.

We believe that this work provides psychologists and developers of future mobile mental health applications with a ready-to-use and effective tool for early diagnosis of mental health issues at scale.

4.2 Problem formulation

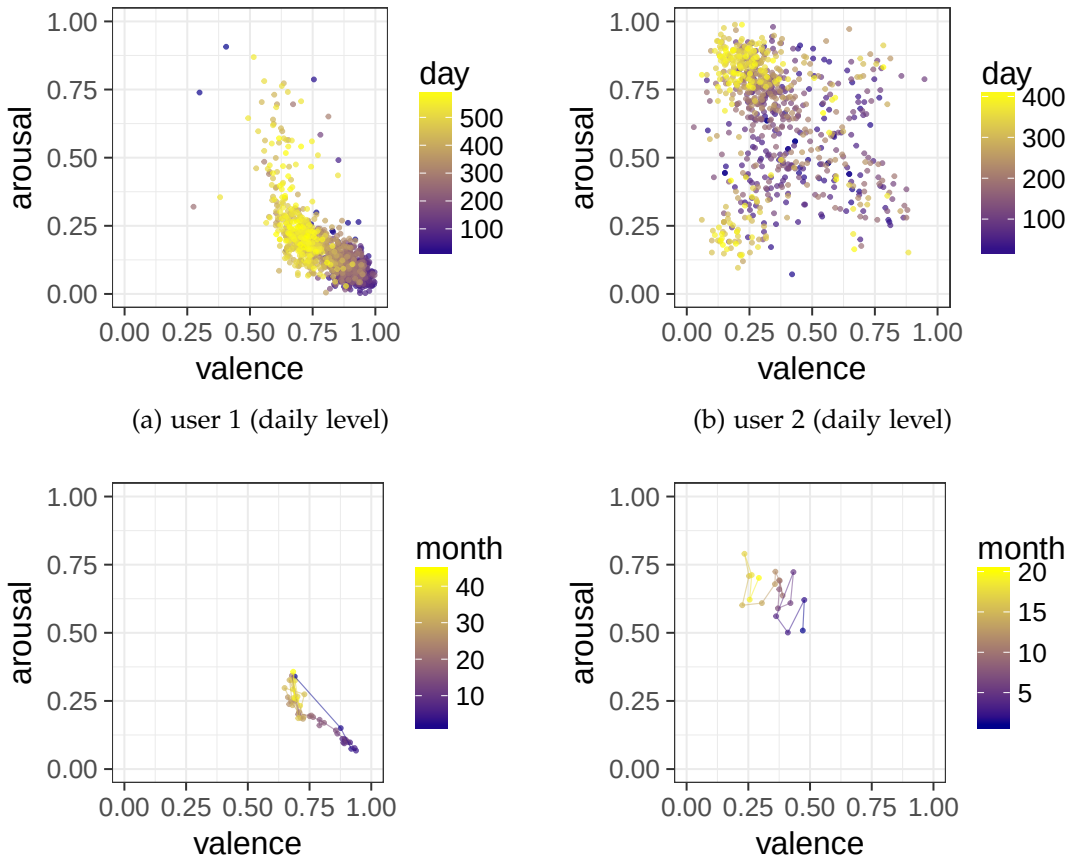
Hypothesis

A machine learning model that employs multi-task learning within an encoder-decoder architecture should be more accurate in forecasting mental health outcomes (fine-grained valence and arousal).

We start by analyzing self-reporting behavior in smartphone applications for mood monitoring. To do so, we consider the EmotionSense dataset (Servia-Rodríguez et al., 2017), as in the previous chapter. For this analysis, we solely consider self-reported mood collected graphically using the Affect Grid scale (Russell et al., 1989). As we discussed in Chapter 3, participants were asked to complete profile-related questionnaires covering a broad range of topics such as demographics, personality and sociability, measured using Likert scales. We will only use such metadata during post-hoc analysis in order to gain insights about model performance at user and group levels.

Sparsity of mood reports. A quick inspection of the dataset revealed that users did not always report even if they were prompted to do so. In the previous chapter, Figure 3.3 showed the complementary cumulative distribution function (CCDF) of moods reported per participant, including those they were prompted to fill *(expected)*, the ones they were prompted to fill but did not *(missed)* and the ones they filled (*complete*). This is also true for users who used the app for large periods. Indeed, those who used the app for 45 or more consecutive days (n=16, 8% of the users) reported, on average, for fewer than half of the expected timepoints. The absence of mood reports might be a symptom of boredom or dissatisfaction with the app, but could also be indicative of mental disorders, especially in cases where users have been reporting anger and depression related feelings.

Variability of mood reports. A longitudinal exploration of the mood reported shows large differences between users in the way they report, in terms of both specific positions on the grid and the area covered. Figure 4.1 shows moods reported by two different individuals who self-reported for at least 300 days, and who are representative of two different behavioral patterns we identified. The first user (user 1 in Fig. 4.1) reports consistently over time, both in the short and long term, and their reports are concentrated on the positive and calm area of the grid. As time elapses, their reports progressively become more negative (but still in the positive area) and active. The second user (user 2 in Figure 4.1) has quite the opposite behavior. That is, at the outset (purple dots), they report mixed affect states during consecutive days (purple dots are almost all over the grid), but, over time, their reports concentrate mainly in the negative and active area.



(a) user 1 (daily level)

(b) user 2 (daily level)

(c) user 1 (monthly average)

(d) user 2 (monthly average)

Figure 4.1: **Longitudinal mood monitoring for 2 illustrative users.** Differences across users in reporting mood levels over time.

Given the longitudinal variability of user's mood, forecasting current and future mood entails different levels of difficulty for different users. For example, it is expected to be much easier to predict for user 1 than for user 2. We will explore this further in §4.5.4.

4.3 Method

We now describe our methodology to build a mood prediction framework capable of handling for the level of noise and sparsity of this kind of data. It consists of a sequence-to-sequence neural network that learns from previous mood sequences to predict future ones (Fig. 4.2). The main advantage of this approach is that, unlike traditional regression, it allows the model to regress to multiple steps into the future by mapping the input sequence to an arbitrary output sequence. The model is composed of an encoder and decoder, each of which are RNNs. The individual units that build up the recurrent networks are *Long Short-Term Memory* units. We use a simplified adaptation of the sequence-to-sequence model proposed for machine translation (Sutskever et al., 2014) as we know exactly how many steps in the future we want to predict, while in translation this length varies (e.g. a sentence in English might have different length in French).

Long Short-Term Memory (LSTM). RNNs are well known to be hard to train especially when employed on sequences with long-term dependencies and patterns (Hochreiter and Schmidhuber, 1997). LSTMs overcome this problem by introducing *memory cells*.

Each LSTM unit has a cell composed of state c_t at time t, also called a memory unit. Sigmoid gates allow the reading and modification of this unit via the input gate i_t, the forget gate f_t, and the output gate o_t. Each unit has four paths, the three gates and the input. At every time-step the unit receives at its four paths inputs coming from two sources: the current mood $\mathbf{x}_t$ and the previous hidden states of all the units in the same layer $\mathbf{h}_{t-1}$. Internally, each gate has another source, the previous cell state c_{t-1}. The inputs are summed along with a bias term b and the total input goes through a sigmoid logistic function. The total input of the input path goes through a non-linearity (*tanh*). The result is multiplied with the activation of the input gate, and then added to the current cell state after multiplying the previous cell state c_{t-1} with the forget gate activation f_t. The final output h_t is calculated by multiplying the output gate o_t with the updated cell state c_t passed through a non-linearity. This happens in a single layer of LSTM units during training (Fig. 4.2). Our encoder and decoder layers are LSTM

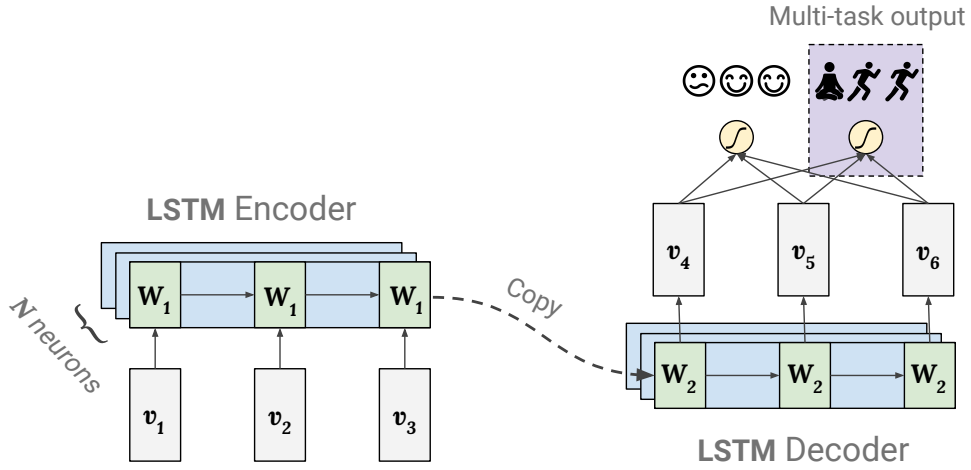


Figure 4.2: **LSTM Encoder-Decoder model**. The mood sequence (v_1,v_2,v_3) passes through an LSTM (states W_1), is transformed to a single vector (dotted) and decoded through another LSTM (W_2) that predicts future mood sequences (v_4,v_5,v_6). Two fully-connected layers are applied to every time-step of the output (yellow circle), one for valence and one for arousal (purple box).

layers like the ones described here. The above updates are summarized as follows:

$$\begin{aligned}
\mathbf{i}_t &= \sigma(W_{xi}\mathbf{x}_t + W_{hi}\mathbf{h}_{t-1} + W_{ci}\mathbf{c}_{t-1} + \mathbf{b}_i) \\
\mathbf{f}_t &= \sigma(W_{xf}\mathbf{x}_t + W_{hf}\mathbf{h}_{t-1} + W_{cf}\mathbf{c}_{t-1} + \mathbf{b}_f) \\
\mathbf{c}_t &= \mathbf{f}_t\mathbf{c}_{t-1} + \mathbf{i}_t tanh(W_{xc}\mathbf{x}_t + W_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c) \\
\mathbf{o}_t &= \sigma(W_{xo}\mathbf{x}_t + W_{ho}\mathbf{h}_{t-1} + W_{co}\mathbf{c}_t + \mathbf{b}_o) \\
\mathbf{h}_t &= \mathbf{o}_t tanh(\mathbf{c}_t)
\end{aligned} \tag{4.1}$$

where $\sigma(\cdot)$ is the sigmoid function, $\mathbf{i}_t$, $\mathbf{f}_t$, and $\mathbf{o}_t$ are the input, forget and output gates, respectively. Since we predict precise mood scores and not binary outcomes, we use the Mean Squared Error (MSE) as the evaluation metric and the loss function to train the model:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{4.2}$$

where Y_i is the vector of n predictions and $\hat{Y}_i$ is the ground truth.

Encoder-Decoder LSTM. The above structure of the LSTM unit outputs the same number of time-steps as the input sequence. Hence, $\mathbf{h}_t$ must connect to additional fully-connected layers to reach the desired dimension of the final output. However, by using simple fully-connected layers we dismiss the sequential nature of the data. To address this, we use a standard LSTM layer as an *Encoder* in order to map the past mood into a fixed length representation with the size of the prediction, and then another LSTM layer as a *Decoder* to reconstruct the original sequence in future steps.

The fixed length representation is feasible through a layer (dotted arrow in Fig. 4.2) designated copy (or *repeat*), which repeats the Encoder 2D output as many times as the output length, in order to create a 3D input for the Decoder. For example, given a week of past mood, we may want to forecast the next two days: the encoder learns to map the past week sequence into a decoded vector of the next two days. A similar model has been applied successfully to video frame prediction, which the authors named *LSTM Future Predictor Encoder-Decoder* (Srivastava et al., 2015).

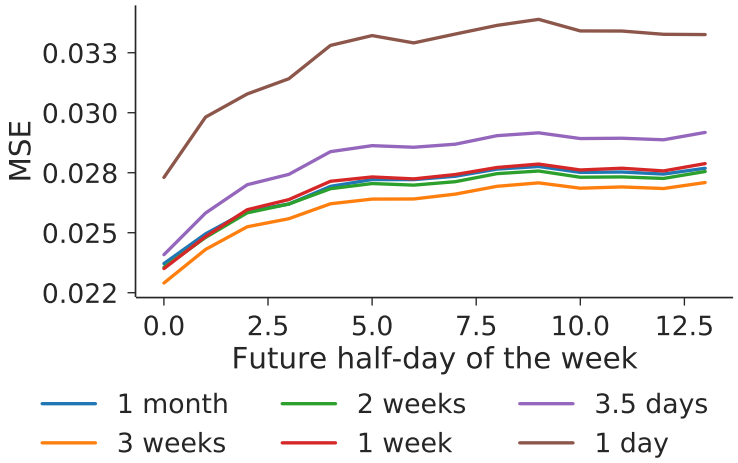
Multi-task Encoder-Decoder LSTM. As we discussed in Chapter 2, multi-task learning is a transfer learning method in which a model learns to predict simultaneously two or more similar tasks. It has been used to reduce overfitting (with *auxiliary targets*), produce better data representations, and in general to improve accuracy in neural networks (Ruder, 2017). Specifically in deep neural networks, this multi-target setup forces the shared weights of the network to optimize both tasks and consequently learn internal representations that reflect on both.

4.4 Evaluation

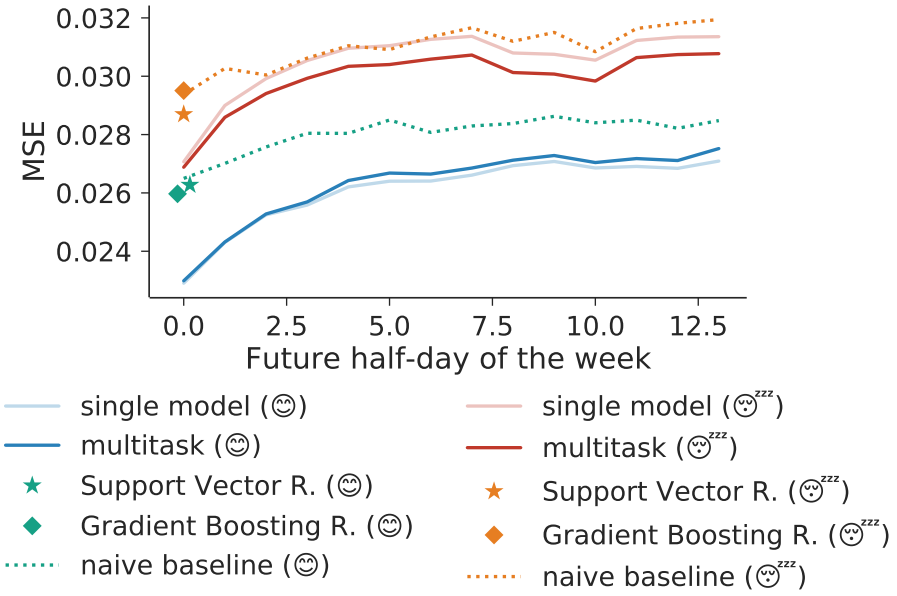
We now evaluate our deep encoder-decoder to forecast mood sequences. We first test different ablations such as a simplified, single-task version of this model and study the optimal length of the input sequence, i.e., number of days in the past, that minimizes the prediction error (§4.5.1). We then explore the performance of multi-task learning for predicting valence and arousal simultaneously (§4.5.2).

Data pre-processing. We selected users who reported more than 100 days ($>$ 200 half-days reports) between May 2013 and October 2016 — the period when the application was most active — resulting in 177,111 unique self-reports from 566 participants. This is the sample we used in our experiments. This subset has similar statistics to the initial sample in terms of valence and arousal: $\mu_{val} = 0.57$ (± 0.17) and $\mu_{aro} = 0.46$ (± 0.19) for the initial dataset, and $\mu_{val} = 0.60$ (± 0.17) and $\mu_{aro} = 0.43$ (± 0.18) for the subset ($\pm$ denotes one standard deviation).

We used a sliding window with step 1 over the mood sequences for each user, obtaining consecutive sequences of 4 weeks of past and 1 week of future moods. We then remove those samples whose *future* moods contained missing values, including these would make training more difficult, resulting in 33,461 final sequences of past and future moods. For the past weeks, we found that only 6, 000 out of 33, 000 (20%) user-weeks had no missing values. Every sequence had on average 15% missing values (i.e. $\mu_{spar} = 6.36$ (± 8.69) for missed time-steps out of 42 steps for a past duration of 3 weeks). For these *past* sequences, we replaced the missing values with zeros. We tested other data imputation methods such as filling with the median of the sequence,



(a)



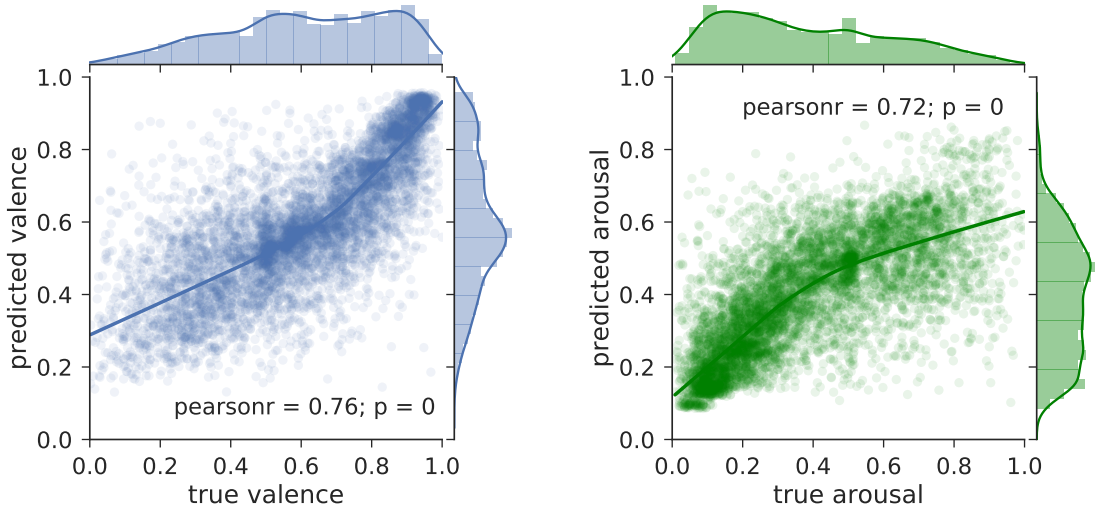
(b)

Figure 4.3: **Model comparison and performance analysis**. (a) How many days into the past should we look for accurate valence prediction? (b) Which is the best model to forecast mood using 3 weeks of past data? (smiley face=valence, sleepy face=arousal)

or min-max scaling to $[0.05, 1]$ but we did not observe any considerable gain on the validation set. To be able to distinguish between real and missing values, we used a Masking layer to skip the missing values during training. In order to prevent overfitting, we split the data into 20% testing and 80% training, ensuring that the users in the test set were completely disjoint, and did not overlap with those in the training set. During training, we used 10% of the training set as validation set to tune our models' hyper-parameters.

Implementation. Our implementation is based on Keras (with Tensorflow backend). We trained two separate models, one for valence and one arousal, for the Encoder-Decoder LSTM model. The input and output data are a matrix $\mathbf{M} \in \mathbb{R}^{s \times t}$, where s are the samples and t the time-steps. After grid-search we found the best-performing number of LSTM units for the Decoder and the Encoder (80 units each). The input layer is a standard Masking layer that skips the time-steps of missing values. In every LSTM layer, a rectified linear unit (ReLU) as well as recurrent dropout of 0.5 probability is applied, to prevent overfitting. The final layer is a standard feedforward neural layer (*Dense*) with a linear activation, that is being applied to every time-step. The objective function minimizes the MSE since this is a regression problem, while the backpropagation optimizer is *Rmsprop*. We train for 300 epochs or until the validation loss stops improving for 15 consecutive epochs. In Appendix A we provide more details regarding the models.

Baselines. We compared our proposed model against a naive baseline based on simply using the average of the past days for predicting future moods (excluding the missing values), a Support Vector Regressor (SVR) and a Gradient Boosting Regressor (GBR). We used the Python's library *sklearn* implementations of an SVR with a radial basis function (RBF) kernel, and a tree-based ensemble model for the GBR, which is reportedly the state-of-art in feature-based machine learning (Olson et al., 2018). SVRs and GBRs do not operate on sequences and assume feature independence, so we extracted 8 representative features from the time-series (non-missing counts, mean, std, min, max, and 25%, 50% and 75% quantiles) and normalized them column-wise to $[0, 1]$. We again exclude the missing values when we calculate those features. We only report the prediction for the *first* future mood since these models cannot regress to sequences.



(a) True and predicted valence.

(b) True and predicted arousal.

Figure 4.4: **Correlations of predictions against ground truth**. Predictive performance of the multi-task model for the first future mood forecast ($p = 0$ denotes a $p < 0.001$).

4.5 Results

4.5.1 How many days should we look back?

Building on previous research (Suhara et al., 2017) we now investigate *how many days we should look back* (i.e., how much history to consider) to predict the sequence of next week's moods. We conduct different experiments to find out which period — 4 weeks, 3 weeks, 2 weeks, 1 week, 3.5 days, or 1 day — predicts with the lowest error the self-reported mood states in the following week. Our assumption is that by using fewer days, the prediction error will increase. By using only the valence axis on the affect grid for training and prediction, we trained a single-task, Encoder-Decoder LSTM model and tested its performance on a test set of disjoint users. Fig. 4.3a shows the MSE for each half-day of the following week for different training sequences. We make the following observations. First, the error increases as we forecast more days into the future. Second, the lowest reported error is $0.022 MSE$ when using 3 weeks of data for training, which corresponds to ± 0.14 error on the affect grid (see Fig. 3.1). Although 1 month includes more time-steps, the length of the optimal sequence of past moods for training is 42 half-days, which corresponds to 3 weeks. This could be attributed either to the inability of the model to learn such long sequences, or that the fourth week of the past does not contain informative and predictive patterns in this dataset. We observe similar behaviors by testing this assumption with our baseline models. Third, our model achieves the highest error when it is trained with just one day of data, i.e., two mood self-reports (± 0.18 error on the affect grid at its worst), followed

by a half-week of data.

4.5.2 How effective are multitask LSTMs?

Motivated by the moderate correlation between valence and arousal (Pearson's r=-0.23) at a significant level (p <0.001), we experiment with learning the two sequences simultaneously in a joint model. Our assumption is that, given this similarity, a multi-task model trained to simultaneously predict valence and arousal would perform better than a single model trained on each separately. To investigate this, we train a single multi-task model with the input and output containing the aligned sequences of valence and arousal in a tensor $\mathbf{T} \in \mathbb{R}^{s \times t \times f}$, where s are the samples, t the time-steps, and f the two features or sequences of mood. The only modification to the single-task model is on the final feed-forward layer, which now has two units, one for each task.

We use the sequence of moods of the previous 3 weeks to predict the sequence of moods in the next week. We do so because in the previous experiment 3 weeks was found to be the period that produces the lowest error in the prediction. From now on, we will refer to these 3 past weeks as *user-weeks*. We use the same data split as in the previous setup and compare different algorithms and approaches. To allow for comparison, we also train a single-task model for the arousal axis using the setup followed earlier with the valence (see 4.5.1), a SVR, a GBR, and a naive baseline that predicts just the average of the past self-reports.

Figure 4.3b shows the MSE for each half-day of the next week for different training sequences and algorithms. Similar to the previous experiment, we observe that the error increases over time. The most interesting result comes from the multi-task learning, which improves the performance of the arousal when trained jointly with the valence, but not the opposite. In general, the arousal axis throughout all of our experiments is more difficult to predict, which reflects on higher errors in all the models. We posit that users might not be as confident evaluating their calmness as they are with their happiness, hence the relationship between the two axes might not be linear. We showed in §5.3 that the heatmap of the two axes forms a *X-shape* (Fig. 3.1). There is evidence that there is a *V-shaped* relationship of arousal as a function of valence (Kuppens et al., 2013). This is in line with previous studies that found that happy/unhappy feelings usually co-occur with higher arousal for some people (reflecting joy/stress), but with lower arousal for others (relaxation/sadness) (Kuppens, 2008).

Regarding the baselines, we observe that the error on the next day's prediction using single-task and multi-task models is lower than those achieved with feature-based algorithms, which fail even to improve the performance of the naive heuristic. In fact,

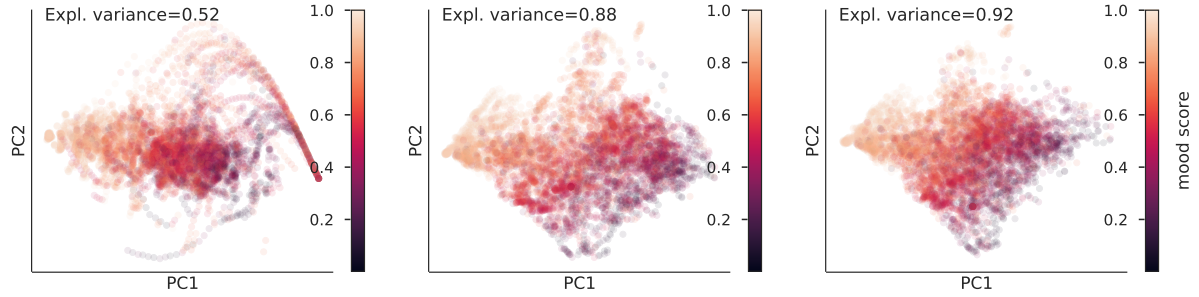
the maximum MSE of 0.032 ($\pm$0.17 on the affect grid), makes them equivalent to using only one day of data for training in the previous experiment (brown line in Figure 4.3a). This motivates the need of using non-linear models like LSTMs. However, and regarding their utility, we believe that simple baselines like these should be encouraged more in time-series forecasting since they provide a fast lower bound. In a more systematic comparison, we compare the error distributions (squared error of predicted and ground truth) of each classifier with a Welch's t-test. Our hypothesis states that multi-task learning will outperform the other classifiers. Indeed, for the valence axis at the first future forecast, the multi-task model presents statistically significant results over the naive baseline ($p < 0.001$), the SVR ($p < 0.001$), and the GBR ($p < 0.001$). Similarly, for the arousal axis, the multi-task model outperforms the naive baseline ($p < 0.05$) and the GBR ($p < 0.05$), and shows a weaker significance against the SVR ($p < 0.10$). For both valence and arousal there is no statistical association between the single-task and multi-task models for the first forecast. However, even if the multi-task models are not better than the single task for the first day, they show lower error during the week for the arousal axis (red lines in Fig. 4.3a). Because of that, we test the forecast of the whole future sequence by taking the median of the week (since the error is not normally distributed) and compare the models. Indeed, the arousal of the multi-task model is significant over the single-task model ($p < 0.05$).

Finally, we inspect the relationship between the predicted and the ground truth scores of valence and arousal for the first future day using our multi-task model (Fig. 4.4). We observe a significant approximation of the two distributions. A non-parametric LOWESS model (locally weighted linear regression) is fitted in order to illustrate the trend. Almost linear trends appear for high valence (happy users) and low arousal (relaxed users), which is also the area with the highest density in the dataset (see Figure 3.1). This is further validated by high and significant ($p < 0.001$) correlations of 0.76 and 0.72 for valence and arousal, respectively.

4.5.3 Understanding the role of the encoder and decoder

We now analyze the role of the LSTM encoder and decoder in predicting sequences of future mood. To do so, we pass the test-set through the multi-task LSTM model and use Principal Component Analysis (PCA) to visualize the response of the network after the encoder and decoder.

Learned representations vs next day's mood. Our test-set is a tensor $\mathbf{T} \in \mathbb{R}^{s \times t \times f}$ where s are the samples, t the time-steps, and f the two features or sequences (for valence and arousal). Fig. 4.5 shows the visualization results for the valence feature and the first time-step in this tensor. Results for arousal and other time-steps follow a



(a) PCA before training

(b) PCA on the LSTM *Encoder* layer activations after training

(c) PCA on the LSTM *Decoder* layer activations after training

Figure 4.5: **Projection of original and latent data**. Visualization of the *Encoder* and *Decoder* responses on the first time-step for the valence axis.

similar pattern, and are omitted here due to space limitations. Data points in the figure are colored according to the first mood to predict (ground truth). We observe that as we move into deeper layers, the network lays out the continuum of positive-negative mood, even though it has been trained to solely predict the next week's mood. Although after the encoder we can already see this continuum, this is more evident after the decoder layer. Apart from qualitative measures, the explained variance of the projections, i.e., the sum of variances of all individual principal components, or more intuitively how much information is lost by going from N to 2 dimensions, increases up to 40% after training (from 0.52 to 0.88 after the encoder, and 0.92 after the decoder).

Learned patterns of individual neurons. We now inspect how the individual neurons of the *decoder* layer *fire* as we pass the test-set through them. Fig. 4.6 shows the mean and standard deviation (denoted in dark and light green respectively) of the activations of the test samples (vertical axis in each subplot) for the 14 time-steps (horizontal axis). We make the following observations: first, the decoder learns various non-linear sequence patterns of future moods, second, some neurons, such as the 4^{th} and 5^{th} in the 6^{th} column, fire almost always with the same exponential decay slope (low deviation), while others, such as the 1^{st}, 3^{rd} and 4^{th} in the second column, are more conservative with almost flat lines (high deviation). Since the decoder is the penultimate layer before the final feed-forward layer that performs the regression, we may interpret it as a proxy for the predictions. For example, one neuron that always fires like the 3^{rd} in the 7^{th} column might be specialized in future mood that rapidly drops and then slowly improves.

4.5.4 Error analysis

We have shown earlier that mood reports might vary within a single user, and especially across a population. Previous research has also found a link between mood variability

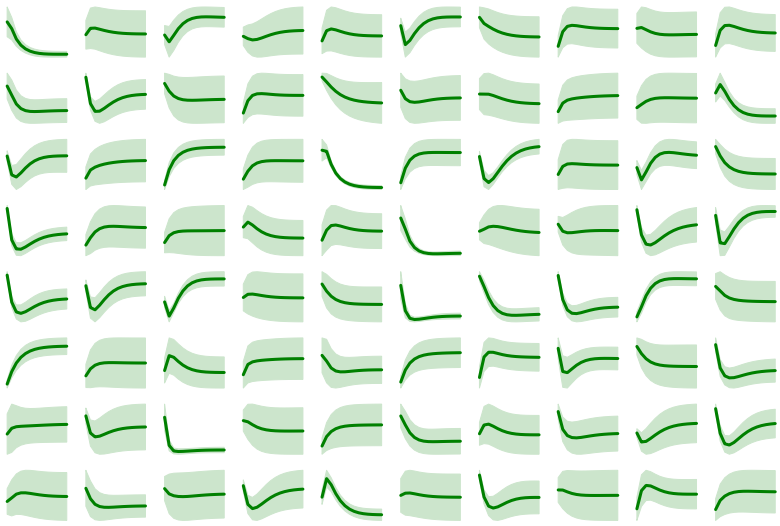


Figure 4.6: **Latent activations**. Visualization of the responses of the 80 neurons of the *Decoder* (one per subplot) for each of the 14 time steps for the valence axis. Light green denotes the mean, and dark green denotes the standard deviation.

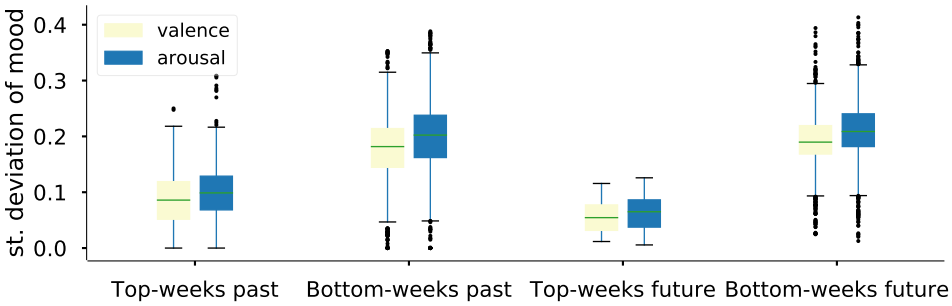


Figure 4.7: **Influence of mood variability on predictive performance**. Deviation of past and future mood by top and bottom performing user-weeks.

and personality traits such as emotional stability (Geukes et al., 2017), and that people tend to exhibit more positive affect on Saturdays than on Mondays (Areni and Burger, 2008). To better understand the performance of our model and assist clinicians in taking informed decisions based on its output, we now investigate how it performs for different mood variability, psychological traits, and days of the week.

To do so, we first average the errors of the predicted sequence on the test-set, obtaining two long tail distributions for valence and arousal (Fig. 4.8). These appear because some user-weeks have errors higher than 0.20 MSE ($\pm$0.44 on the affect grid), although the majority of the distribution resides below 0.025 MSE. Indeed, for valence, more than 10% of the user-weeks have MSE close to zero. We then divide the MSE-distributions in three equally-sized samples, and consider the 1^{st} quantile as the top-performing user-weeks, and the 3^{rd} quantile as the worst-performing.

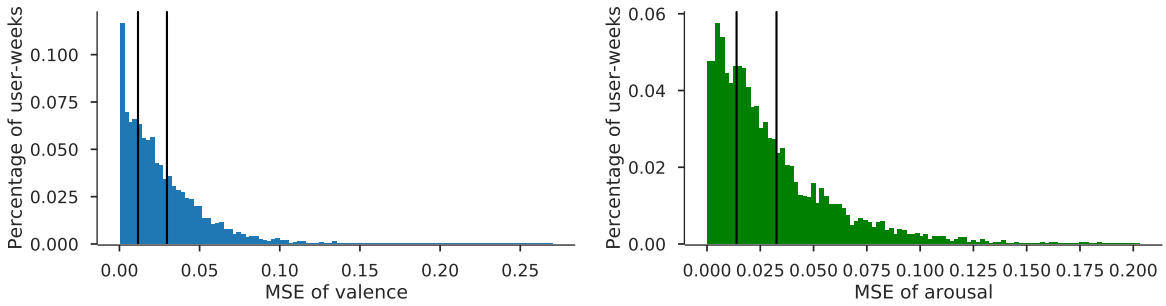


Figure 4.8: **Error distributions across the two mood dimensions**. Distribution of *avg* MSE for valence and arousal. The MSE corresponds to the average predictions of all future days of the week for the multi-task model. Black bars denote the 1^{st} and 3^{rd} quantiles (at 33% and 66%, respectively).

Mood variability

We first investigate the influence of the mood variability in the best and worst performing user-weeks. We assess the mood variability for each user-week in these two groups by computing the standard deviation of both the mood of the three past weeks (ignoring the missing values) and the mood of the future week. The boxplot in Fig. 4.7 shows the difference between these two groups. We observe that (i) the MSE increases with the variability of the (past or future) mood, and (ii) valence and arousal have similar median deviation, although the median of the arousal is slightly higher. The lowest deviation is on the future top-weeks, where there are no outliers in the boxplot, which means that the model is very reliable for those user-weeks with more stable future mood. Finally, the absolute mood differs between the bottom ($\approx$ 0.2 std) and top quantiles ($\approx$ 0.1 std) of the error. Specifically, the model is more reliable for user-weeks with high valence and low arousal as we saw in Fig. 4.4.

Personality traits

We now study the influence of personality traits in the best and worst performing user-weeks. We consider those individuals with samples in the 1^{st} and 3^{rd} quantile of the prediction error distributions (Fig. 4.8) who completed the personality questionnaire. This includes questions regarding the *Big-Five* personality traits (Gosling et al., 2003): Agreeableness, Conscientiousness, Emotional Stability, Openness, and Extraversion, answered through a discrete Likert scale with values normalized in [0,1]. Note that not every user completed the personality questionnaire. Thus, even though each original quantile contains the same number of user-weeks (2188), our sample shrinks to 701 (1^{st}) and 1082 (3^{rd}) user-weeks for valence when we consider only users who responded the questionnaire, and to 687 (1^{st}) and 1207 (3^{rd}) user-weeks for arousal (some users might appear in both quantiles).

Table 4.1: **Personality and model performance**. Differences on personality between the top and bottom quantiles, broken down by valence and arousal. Significance is represented with * = $p < 0.05$, ** 0.01, *** 0.001.

	Valence				Arousal			
	Mean				Mean			
	Top	Bottom	t-stat	Sig	Top	Bottom	t-stat	Sig
Agreeableness	0.67	0.61	6.58	***	0.68	0.60	9.53	***
Conscientiousness	0.66	0.59	5.55	***	0.70	0.64	4.72	***
Emotional Stability	0.33	0.25	6.42	***	0.38	0.25	11.89	***
Openness	0.77	0.61	16.48	***	0.78	0.66	12.55	***
Extraversion	0.24	0.29	-4.00	***	0.32	0.27	3.62	***

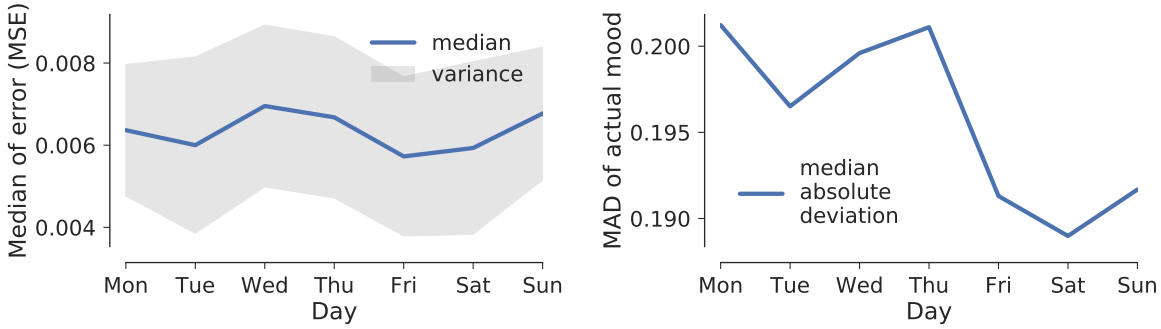
Table 4.2: **Personality and correlation with model's error.** Pearson's correlation (r) of the prediction error (MSE) with the personality of top and bottom quantiles, broken down by valence-arousal. Significance is represented with * = p ¡ 0.05, ** 0.01, *** 0.001.

	Valence				Arousal			
	r with MSE				r with MSE			
	Top	Sig	Bottom	Sig	Top	Sig	Bottom	Sig
Agreeableness	-0.12	**	-0.06	*	0.01		-0.10	***
Conscientiousness	-0.06		0.03		-0.20	***	0.29	***
Emotional Stability	-0.19	***	-0.00		-0.11	*	0.00	
Openness	-0.03		-0.07	*	0.03		0.14	***
Extraversion	-0.06		-0.00		0.11	*	-0.06	*

We first perform a Welch's t-test to check whether there are significant differences between the personality traits of the users in the two quantiles (Table 4.1). Significant differences were found for all traits, with special relevance for Openness. That is, users for whom the model forecasts happiness *and* calmness more accurately tend to be more open to new ideas and showcase creativity, intellectual curiosity, and a preference for novelty.

Previous research found that Emotional Stability, Extraversion, Agreeableness, and sometimes Conscientiousness were related to decreased variability in affect (Geukes et al., 2017). Earlier, we showed that users in the top and the bottom performing quantiles differ in terms of their mood stability, while here we see also that all of their personality traits are significantly related with the performance of the model. In our case, higher Openness might be associated with the nature of our experiment and data collection since users who are more open to new technologies might use the app more honestly and therefore becoming more predictable.

We finally check whether increments in personality scores increase or decrease the error. Table 4.2 shows the correlation of the error with the personality traits. For valence users in the top quantile we observe that our model is increasingly more



(a) (b)

Figure 4.9: **Day and model performance**. Contribution of the day of the week to the median error for the first future mood (valence) (a). Comparison with the actual mood variability (b).

accurate for Emotionally Stable users (r = -0.19). We do not observe the reverse effect in the bottom quantile. For arousal, the model is more accurate for users with high Conscientiousness — which is associated with self-discipline — (r = -0.20) and a reverse effect appears on the bottom quantile (r = 0.29).

Day of the week

We now investigate the impact of the day of the week on the accuracy of our model. We consider the error of the first mood in the sequence of predicted moods, grouped by the day of the week. The distribution of this error is similar to the distribution of the *average* error of the sequence of predicted moods in Fig. 5.3, but skewed towards lowest errors since our errors are lower for tomorrow's mood (first day in the sequence). We group and average (i) the errors by the day of the week, and (ii) the actual mood of this day across all the user-weeks in the test set. We observe that the distributions of the actual mood (Fig. 3.1) and the error (Fig. 5.3) across all user-weeks in the test set are very different. While the actual mood has a bimodal shape, the error resembles more a long tail. Thus, since we compare distributions with non-uniform shapes we use robust statistics, such as the median or median absolute deviation (the median of the absolute deviations from the data's median: $MAD = median(|X_i - median(X)|)$).

We obtain the median and variance of the error for each day of the week, and the median absolute deviation (MAD) of the actual mood for each day. Fig. 4.9 shows the median of the MSE (a) and the MAD of the actual mood (b) across different days of the week. We observe that our model (Fig. 4.9 (a)) is more accurate on Tuesdays, Fridays and Saturdays, while the highest median error is on Wednesdays. This is consistent with the trend observed on the variability of the actual mood (Fig. 4.9 (b)), where on Fridays and Saturdays there were fewer differences across the reported moods.

Error analysis implications

Mood variability, personality and day of the week could play a role in the performance of our model. Clinicians may wish to *screen* their patients with brief personality questionnaires to assess the reliability of the sequences of moods predicted. For example, Openness affects performance in both dimensions, whereas Emotional Stability affects valence, and Conscientiousness arousal. Clinicians should also consider the day of the week when forecasting sequences of mood, however, we think that this, due to the high variance, needs to be validated by external datasets.

We also acknowledge the caveats of our model. The average error (MSE) of our model is low across the population (± 0.14 error on the affect grid valence for valence, ± 0.16 for arousal). However, the fact that it performs better for Emotionally Stable users, and users with low mood variability, might limit its utility in patients with mental disorders. Further analysis of the trajectories of mood reported by *unstable* individuals is required to build accurate models for this specific population. Moreover, the studied outcome should be affected by a variety of environmental and genetic factors and additional data collected in this study could improve forecasting. We leave this for future work.

4.6 Discussion

Most of the related works focus on binary outcomes such as depression prediction. In particular, the *Deepmood* (Suhara et al., 2017) study analyzed 2,382 users over 2 years. In contrast to this work, our model does not aim to distinguish between healthy and depressed patients, but to predict a sequence of real-valued moods. Binary prediction is ubiquitous in the mood prediction literature, where mood is simplified to a binary state (Taylor et al., 2017; Servia-Rodríguez et al., 2017), and extreme depression is considered in the same class as moderate unhappiness. Since neutral mood might be uninformative and make the predictions harder, authors often omit the middle 40-60% of reports. Instead, we use regression to predict precise mood scores.

4.7 Conclusion

This chapter introduces a new end-to-end, stand-alone ML model to forecast future sequences of mood from previous self-reported mood. Contrary to previous research on classifying between extremes of mood using data collected in controlled experiments with limited number of participants, we forecast exact values of valence and arousal from noisy and sparse reports collected in the wild.

Experiments using a real-world dataset revealed that (i) 3 weeks of sparsely reported mood is the optimal number to accurately forecast mood, (ii) multi-task learning learns both dimensions of mood — valence and arousal — with higher accuracy than when training separate models, and (iii) mood variability, personality traits and day of the week play a key role in the performance of our model. We believe that this work provides psychologists and developers of future mobile mental health applications with a ready-to-use and effective tool for early diagnosis of mood issues at scale.

This second empirical chapter of the thesis builds on the topic introduced by the previous one by modeling mobile-measured mood. However, it slightly diverts from the general theme of the thesis by using only self-reported mobile surveys, instead of passively sensed data. While clustered or binary mood outcomes were reasonably predictable as we saw in the previous chapter, we observed that adding sensor data was not helpful in fine-grained forecasting of mood. Therefore, in this chapter we focused on modeling the sequences of mood reports, motivated by the — as of now — limited understanding of the temporal dynamics of mood variability of individuals. In these two chapters (3 and 4), we focused on single-purpose models tailored to mental health tasks. In the following chapter, we take a more generalized approach on using mobile sensor data for learning latent features which can be used in many health-related tasks. In particular, we develop self-supervised models which learn a physiological representation that can transfer to many tasks pertinent to physical health.

Chapter 5

Learning generalizable physiological representations with self-supervision

Data without generalization is just gossip
–Robert M. Pirsig

5.1 Introduction

In the previous chapters we proposed models which improve mood prediction in the wild. We now switch focus to physical health and in particular using wearables which provide behavioral and physiological data for population-health inferences. We present a self-supervised model which exploits the multimodal data of modern wearables to learn meaningful representations which generalize to several outcomes with transfer learning. These results also motivate the use of free-living data for more accurate prediction of cardio-respiratory fitness, which is further explored in Chapter 6.

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. In particular, even though deep learning has shown great promise in human activity recognition (HAR) tasks using wearable sensor data (Yang et al., 2015; Ma et al., 2019; Alsheikh et al., 2015), it relies, by and large, on purely labeled datasets which are costly to collect (Bulling et al., 2014). In addition, they are obtained in laboratory settings and hence might not generalize to free-living conditions where behaviors are more diverse, covering a wide distribution of activities (Krishnan et al., 2018). Unsupervised learning is a qualified candidate to solve this label scarcity problem in wearable data, particularly given the vast amounts that can be collected in free-living

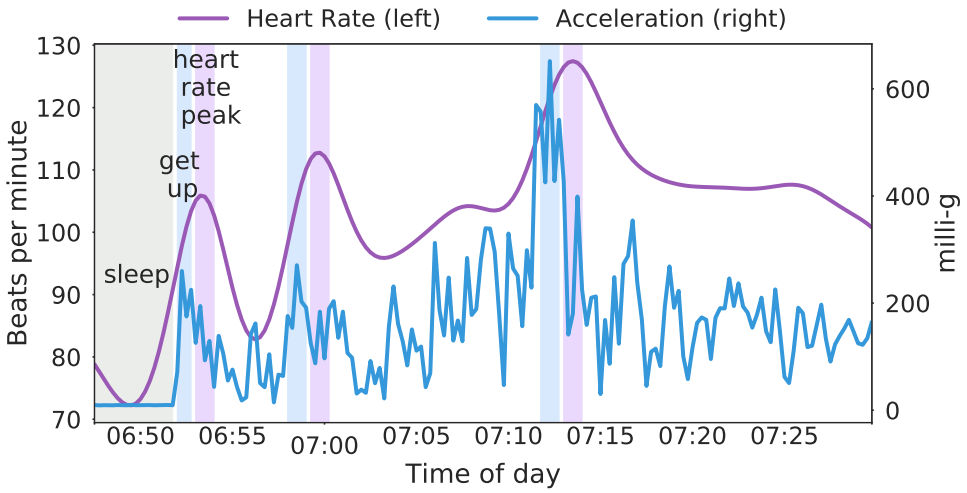


Figure 5.1: **Heart rate and acceleration temporal dynamics**. Illustrative visualization of the relationship between movement and heart rate responses (randomly selected participant). Shaded areas show this lagging relationship.

conditions. Recent models have effectively utilized unlabeled activity data to learn useful summary representations of sensor signals (Aggarwal et al., 2019). Notwithstanding 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. 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 (this relationship is conceptualized in Figure 5.1). Multimodal learning has proven beneficial in supervised tasks such as fusing images with text to improve word embeddings (Mao et al., 2016), video with audio for speech classification (Ngiam et al., 2011), or different sensor signals for HAR (Radu et al., 2018). 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 (Owens et al., 2016) or, in our case, activity to heart rate), using unlabeled data with supervised objectives (Lan et al., 2020). 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 >280, 000 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 rep-

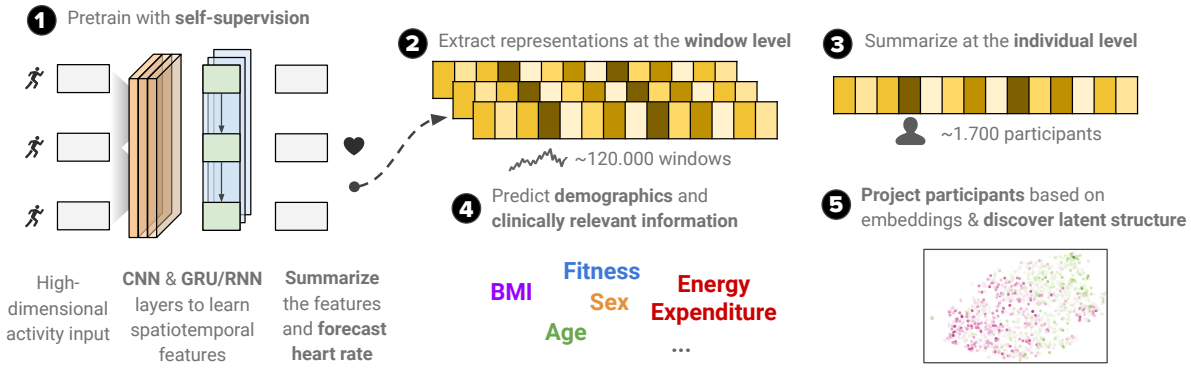


Figure 5.2: **Schematic of model architecture and tasks**. (1) Self-supervised pre-training, CNN + GRU temporal model for HR forecasting, (2) window-level representation extraction, (3) aggregation and summary at the individual level, (4) inference of health and fitness related outcomes, and (5) clustering through PCA based on embedding space and latent structure discovery.

resentations (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 example, they can be used to predict variables associated with individuals' health, fitness and demographic characteristics (AUC >70), outperforming unsupervised autoencoders and common biomarkers. Overall, we propose a multimodal self-supervised method[1] for behavioral and physiological data with implications for large-scale health and lifestyle monitoring.

This chapter 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.

[1]Code and sample data: https://github.com/sdimi/Step2heart

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 biomarkers. 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 behaviors 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 5.2 and our code/models are publicly available.

5.2 Method

Hypothesis

A machine learning model that is trained with a self-supervised objective — using physical activity data — should generalize better in multiple downstream health-related tasks.

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 5.1: **Notation.**

Notation	Description
$\mathcal{D}_{train}$, $\mathcal{D}_{test}$	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
$\tilde{\mathcal{D}}_{train}$, $\tilde{\mathcal{D}}_{test}$	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 5.1. 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 (Spathis et al., 2019) and reinforcement learning for video prediction (Ha and Schmidhuber, 2018). 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 (Sanchez-Lengeling et al., 2019)–, 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 (Le and Mikolov, 2014), we develop a simple aggregation method of sensor windows to the user level. This is a common issue in the literature (Chen et al., 2019).

5.2.1 Model architecture

As shown in Figure 5.2 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* (Yao et al., 2017), 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 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.

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

$$C(s) = (\mathbf{x} * f)(s) = \sum_{i=0}^{k-1} f(i) \cdot \mathbf{x}_{s-i} \tag{5.1}$$

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 $\mathbf{X}$.

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) (Cho et al., 2014). 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:

$$\begin{aligned}
\mathbf{r}_t &= \sigma(W_r \mathbf{x}_t + U_r \mathbf{h}_{t-1} + \mathbf{b}_r) \\
\mathbf{z}_t &= \sigma(W_z \mathbf{x}_t + U_z \mathbf{h}_{t-1} + \mathbf{b}_z) \\
\tilde{\mathbf{h}}_t &= tanh(W \mathbf{x}_t + U(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}) \\
\mathbf{h}_t &= (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t
\end{aligned} \tag{5.2}$$

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 : $\mathbf{X}$ (sensors), $\mathbf{M}$ (metadata), $\mathbf{y}$ (target HR)
Output: $\tilde{\mathbf{E}}$ (user-level embedding), $\tilde{\mathbf{y}}$ (target variable)
while *neural network* θ *not converged* **do**
  pass $\mathbf{X}$ through CNN/RNN layers (eq. 5.1 & 5.2);
  pass $\mathbf{M}$ through reLU layers;
  concatenate outputs in $\mathbf{E}$;
  forecast & backpropagate with joint loss $\mathcal{L}$ (eq. 5.5);
end
use trained network θ to extract embeddings $\mathbf{E}$;
aggregate $\mathbf{E}$ to the user-level $\tilde{\mathbf{E}}$ with average pooling;
train a linear model to predict target variables $\tilde{\mathbf{y}}$;

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.

5.2.2 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) (Tanaka et al., 2001) 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 overestimates.

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

$$\mathcal{L}(\xi_i|\alpha) = \begin{cases} \alpha\xi_i & \text{if } \xi_i \geq 0, \\ (\alpha - 1)\xi_i & \text{if } \xi_i < 0. \end{cases} \tag{5.3}$$

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. (5.3) 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:

$$\mathcal{L}_Q(\mathbf{y}, \mathbf{f}|\alpha) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(y_i - f(\mathbf{x}_i)|\alpha) \tag{5.4}$$

The quantile loss (or tilted/pinball loss in the literature) can be seen as a *tilted* version of the l_1 loss which estimates the unconditional median. Instead, if a prediction falls below a given quantile (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 (Rodrigues and Pereira, 2018) as well as reinforcement learning (Dabney et al., 2018).

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:

$$\begin{aligned}\mathcal{L}_{MSE+Q} = \frac{1}{N}\sum_{i=1}^{N}\Bigg((y_i - f(\mathbf{x}_i))^2 \\ + \sum_{j=1}^{J} max\Big(\alpha_j(y_i - f(\mathbf{x}_i)^{(\alpha_j)}), \\ (\alpha_j - 1)(y_i - f(\mathbf{x}_i)^{(\alpha_j)})\Big)\Bigg)\end{aligned} \tag{5.5}$$

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 5.3 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 very 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 the next sections, we apply different ablations of these terms to evaluate their impact.

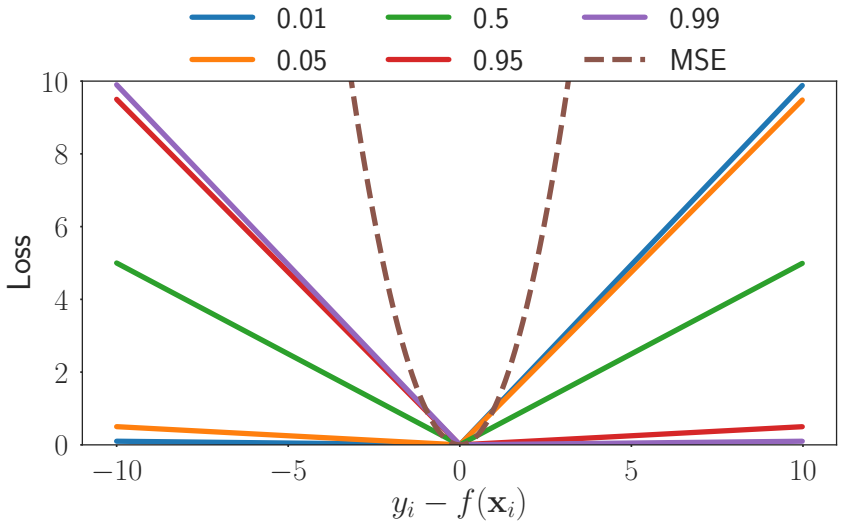


Figure 5.3: **Quantile vs MSE loss**. Illustration of the relationship between the prediction and the loss with respect to the shapes of the MSE and various levels α of quantiles. Simulated data, the true value is $y_i = 0$.

5.3 Evaluation

Data Pre-processing. We used the *Fenland* dataset we introduced in Chapter 2. All participant heart rate data collected during free-living conditions underwent pre-processing for noise removal (Stegle et al., 2008). Similarly, all accelerometer data was auto-calibrated to local gravity, the 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 of day and seasonality can have a significant impact on physical activity, such as sleeping and commuting, we encoded the sensor timestamps using *cyclical temporal features* T_f (Chakraborty and Elzarka, 2019). Here, we encoded the month of the year and the hour of the day as (x, y) coordinates on a circle:

$$T_{f_1} = sin\Big(\frac{2 * \pi * t}{max(t)}\Big) \qquad (5.6) \qquad\qquad T_{f_2} = cos\Big(\frac{2 * \pi * t}{max(t)}\Big) \qquad (5.7)$$

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

Table 5.2: **Data description**. *Seq.* denotes sequential measurements (timeseries), while *Inp.* the inputs to the pre-training task. (◇ feature used in some models, see Results)

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

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 5.2 (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

Table 5.3: **Forecasting task results.** Ablation test to compare the HR forecasting error using different input modalities and baselines.

	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)

using the Adam (Kingma and Ba, 2014) optimizer for 300 epochs or until the validation loss stopped improving for 5 consecutive epochs [2]. 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. We provide more details about the models in the Appendix A.

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[3]. 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.

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

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

Table 5.4: **Loss function results.** Ablation test to compare the best performing model with regards to different loss functions.

	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)

5.3.1 Baselines and metrics

For our baselines, we used naive lower bounds as well as modern ML models (similar to those used in previous works (Ni et al., 2019; Hallgrímsson et al., 2018)):

- **Convolutional Autoencoder:** A convolutional autoencoder learns to compress the input data ($\mathbf{X} \rightarrow \mathbf{X}$) with a reconstruction loss. This unimodal baseline uses movement data only and is conceptually similar, albeit simpler, to (Aggarwal et al., 2019; Saeed et al., 2019). 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.

- **Gradient Boosting (XGboost)**: gradient boosting machines are among the best performing ML methods (Chen and Guestrin, 2016). 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 at any one time 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 $\mathbf{X}$ 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).

5.4 Results

5.4.1 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 5.3. *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 biomarker (Fox et al., 2007).

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 (Jaques et al., 2017).

Impact of the joint loss. When comparing different loss functions with the best performing model *Step2Heart*(A/R/T), we see (Table 5.4) 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 select the former as the one with the lowest average error.

Table 5.5: **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 $\times 100$ for better legibility. (*percentage of explained variance by compressing the dimensionality of embeddings with PCA)

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			

5.4.2 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 5.5.

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 (Hallgrímsson et al., 2018) 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 (Manini et al., 2006). Similarly, VO_2max is prospectively associated with the incidence of type 2 diabetes (Katzmarzyk et al., 2005).

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 (Hinton et al., 2015).

Visualizing the latent space. Qualitatively, we visualized the resulting *latent* space in 2D with t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008) as shown in Figure 6.6. 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 behaviors (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).

5.5 Discussion

Our results showcase the generalizability of the proposed models to solve different tasks pertinent to physiological and behavioral data. Most studies in wearable and mobile sensing have been focused on human activity recognition using mobile devices (Saeed et al., 2019; Tang et al., 2021) and emotion recognition using ECG data (Sarkar and Etemad, 2019), 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 (Hallgrímsson et al., 2018). However, this is an auto-encoder based approach requiring a historical input of one 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 (Chen et al., 2021), while abstract (non-sensor related) attributes such as gender and age have been proved to be predictable with wearable embeddings (Wu et al., 2020). 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 (McConville et al., 2018).

Viewing the upstream task (HR forecasting) in isolation, we consider the error acceptable for real-world deployments, especially in cases of energy-constraint environments where the heart rate sensors could be precluded (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 this may affect the average error).

Further, some interesting extensions to the transfer learning experiments include 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 (Doherty et al., 2017) or the *All of Us* (of Us Research Program Investigators, 2019) 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), and pioneers in representation learning (Shin et al., 2016), 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 (Chen et al., 2021), 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.

Regarding broader implications, we should acknowledge that 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 mod-

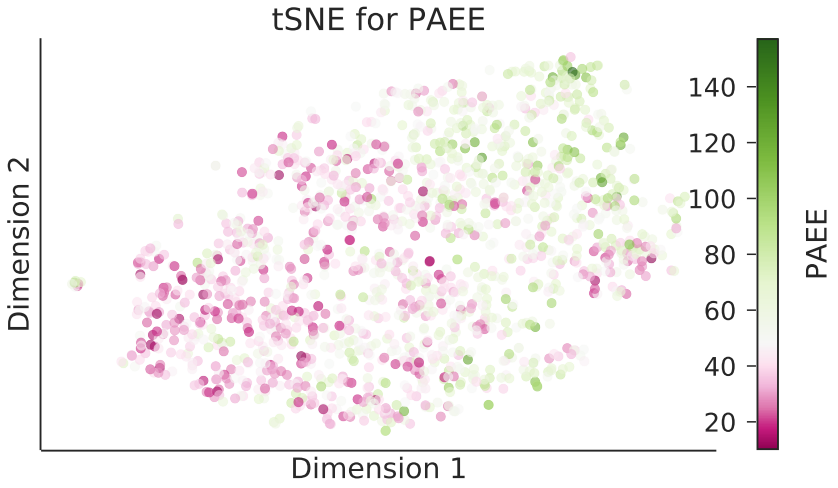


Figure 5.4: **Model embeddings for transfer learning visualized with t-SNE.** 2D representation of the test set colored with the PAEE outcome, with the colorbar showing the extreme values (the median participant has PAEE=48, white color). See Table 5.5 for full results.

els for pattern recognition is impractical. Through self-supervised learning, we can leverage this unlabeled 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, which is of the 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 with *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.

5.6 Conclusion

In this Chapter, 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 biomarkers. 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 (Chen et al., 2020; Tang et al., 2020) 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.

This third empirical chapter of the thesis builds on the two previous chapters by employing machine learning on mobile-measured data. While the previous chapters focused on smartphone-based assessment of mental health, here we investigated wearable-based inferences of outcomes related to physical activity and metabolic health. We believe that there are many associations between physical and mental health as evidenced by the improvement of the mood prediction when using movement data in Chapter 3. Therefore, models developed for mental health outcomes could be also used for physical activity outcomes and vice versa. For instance, the upstream model of this chapter is an encoder-decoder model conceptually similar to that of Chapter 4; the former is used as the main model since we are interested in the task of mood forecasting while the latter is used simply as a feature extractor. In the end of the day, the problem formulation defines the task and the modeling; for example, continuous heart rate sensing is impossible with smartphones and hence it can only be leveraged with a wearable dataset.

While this chapter showed how generic models could transfer to many coarse-grained outcomes related to one's physiology, it also highlighted the promise of wearable-based estimation of cardio-respiratory fitness, a well-established predictor of metabolic disease and mortality. The following chapter expands this line of work with models tailored to fine-grained fitness prediction through extensive experiments on a larger dataset of the *Fenland* Study.

Chapter 6

Longitudinal fitness prediction with wearables

Τίς εὐδαίμων, ὁ τὸ μὲν σῶμα ὑγιής, τὴν δὲ ψυχὴν εὔπορος, τὴν δὲ φύσιν εὐπαίδευτος [1]
–Thales of Miletus

6.1 Introduction

In the previous chapter, we introduced a new self-supervised task which sets the heart rate responses to activity as a supervisory signal and showed promising results across a range of binary outcomes pertinent to one's physiology. In this chapter, we delve into the task of fitness prediction. We use a larger set of the *Fenland* Study than that of Chapter 5, which includes only data from the wearable-ECG, and most importantly, a longitudinal cohort who repeated the protocol (both free-living and lab-based) almost a decade later. The latter allows us to validate the adaptability of our models in predicting long-term fitness change. Our findings motivate the complementary use of wearables with deep learning models for fine-grained and more accurate fitness estimation, compared to established estimations of commercial wearables.

Cardiorespiratory fitness is a well-established predictor of metabolic disease and mortality. Fitness is directly measured as maximal oxygen consumption (VO_2max), or indirectly assessed using heart rate response to a standard exercise test. However, such exercise testing is costly and burdensome, limiting its utility in healthcare and large-scale population studies. Fitness can also be approximated using resting heart rate and self-reported exercise habits, but accuracy is low compared to estimates based on dynamic data.

Given these limitations, VO_2max is usually not assessed directly in most settings.

[1] Who is happy? One who has a healthy body, a resourceful mind and a docile nature.

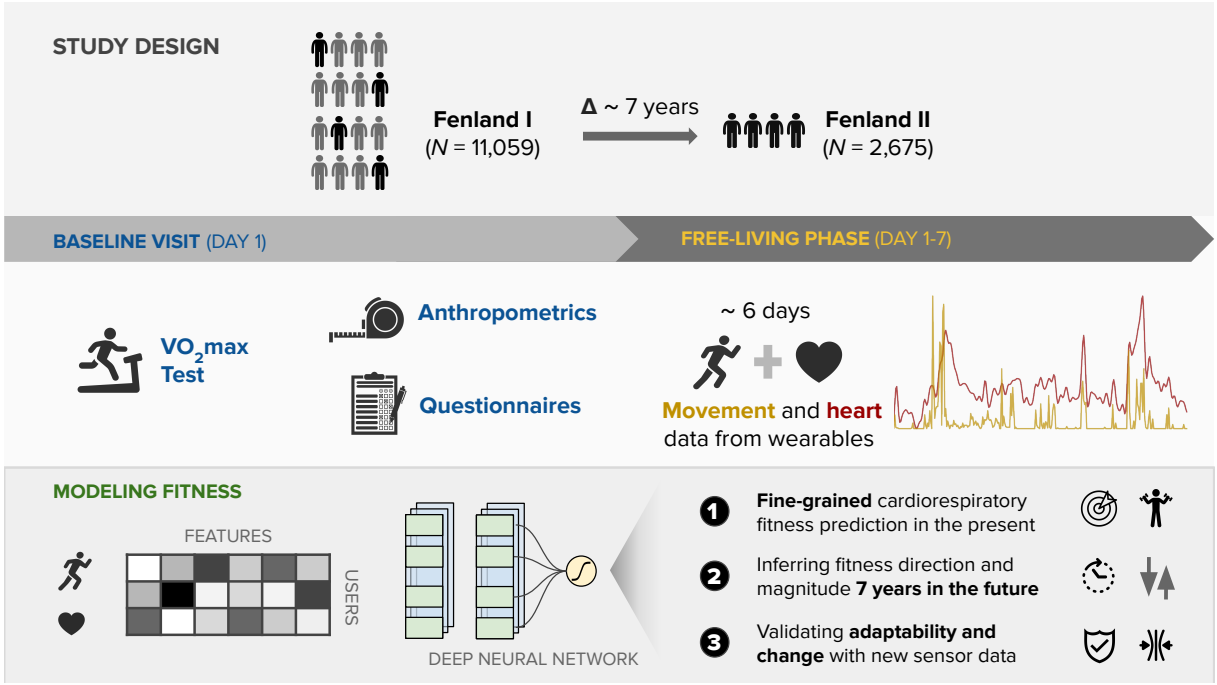


Figure 6.1: **Study and experimental design**. We include 11, 059 participants with treadmill and wearable sensor data at baseline (*Fenland* I, 2005-2015) and a longitudinal subsample of 2, 675 participants who were retested approximately 7 years later (*Fenland* II). Following this baseline clinic visit, participants were fitted with a combined heart rate and movement sensing device which they wore during free-living conditions for approximately six days.

Instead, indirect methods to estimate VO_2max through submaximal exercise tests have been developed, aiming to predict VO_2max from the heart's response to incremental exercise at submaximal intensities (Abut et al., 2016; Brage et al., 2007). These models are based on the parallel increase of heart rate and VO_2 consumption during exercise, and assume a linear relationship between work rate and heart rate that holds at maximal intensities (Davies, 1968; Tanaka et al., 2001). Although some studies have measured submaximal VO_2 (Jurca et al., 2005; Wareham et al., 1997; Dunbar, 1992), the most common approach is to estimate the oxygen cost of standardised work loads and only directly measure the heart rate response (Brage et al., 2007; Weller et al., 1995; Christensen et al., 2012; Assah et al., 2015). Even though these submaximal tests are valuable alternatives to maximal exertion tests, particularly for older and non-athlete populations, they still require standardised testing and access to ergometry equipment, thus limiting their applicability in large-scale population studies due to scalability, cost, time consumption and potential risks associated to exertion (Noonan and Dean, 2000).

Modern wearables capture dynamic heart rate data which could improve fitness prediction. In this work, we analyze movement and heart rate signals from wearable sensors in free-living conditions from 11,059 participants who also underwent a standard exercise test. We develop a deep neural network model that leverages sensor information to predict maximal oxygen uptake (VO_2max), yielding high correlation (r = 0.82, 95% CI 0.80-0.83), when compared to the ground truth in a holdout sample. This model outperforms conventional non-exercise fitness models and traditional bio-markers using measurements of normal daily living without the need to undertake a specific exercise test. Additionally, we show the adaptability and applicability of this approach for detecting fitness change over time in a longitudinal subsample (n = 2,675) who repeated measurements after 7 years. The latent representations that arise from this model pave the way for fitness-aware monitoring and interventions at scale.

This work makes the following contributions:

- We analyze movement and heart rate signals from wearable sensors in free-living conditions from 11, 059 participants who also underwent a standard exercise test.
- We develop a deep neural network model that leverages sensor information to predict maximal oxygen uptake (VO_2max), yielding high correlation (r = 0.82, 95% CI 0.80-0.83), when compared to the ground truth in a holdout sample.
- This model outperforms conventional non-exercise fitness models and traditional biomarkers using measurements of normal daily living without the need to undertake a specific exercise test.
- Additionally, we show the adaptability and applicability of this approach for

detecting fitness change over time in a longitudinal subsample (n = *2, 675*) who repeated measurements after seven years. The latent representations that arise from this model pave the way for fitness-aware monitoring and interventions at scale.

6.2 Method

Hypothesis

A machine learning model that is trained on a combination of free-living wearable data and traditional biomarkers should be more accurate in predicting lab-measured cardio-respiratory fitness.

6.2.1 Cardiorespiratory fitness assessment

VO_2max was predicted in study participants using a previously validated submaximal treadmill test (Gonzales et al., 2020). Participants exercised while treadmill grade and speed were progressively increased across several stages of level walking, inclined walking, and level running. The test was terminated if one of the following criteria were met: 1) the participant wanted to stop, 2) the participant reached 90% of age-predicted maximal heart rate (208-0.7*age) (Tanaka et al., 2001), or 3) the participants exercised at or above 80% of age-predicted maximal heart rate for 2 minutes.

6.2.2 Free-living wearable sensor data processing

Participants were excluded from this analysis if they had less than 72 hours of concurrent wear data (three full days of recording) or insufficient individual calibration data (treadmill test-based data). All heart rate data collected during free-living conditions underwent pre-processing for noise filtering (Stegle et al., 2008). Non-wear detection procedures were applied and any of those non-wear periods were excluded from the analyses (Brage et al., 2015). This algorithm detected extended periods of non-physiological heart rate concomitantly with extended ($>$ 90 minutes) periods that also registered no movement through the device's accelerometer. We converted movement intensities into standard metabolic equivalent units (METs), through the conversion 1 MET = 71 J/min/kg (3.5 ml $O_2 \cdot$ $min^{-1} \cdot$ kg^{-1}). These conversions where then used to determine intensity levels with $\leq$ 1.5METs classified as sedentary behavior, activities between 3 and 6 METs were classified as moderate to vigorous physical activity (MVPA) and those $>$ 6METs were classified as vigorous physical activity (VPA).

Since the season can have a big impact on physical activity in terms of how it affects workouts, sleeping patterns, and commuting patterns, we encoded the sensor timestamps using *cyclical temporal features* T_f. Here we encoded the month of the year as (x, y) coordinates on a circle as previously shown in Chapter 5. The intuition behind this encoding is that the model will "see" that for example December (12th) and January (1st) are 1 month apart (not 11). Considering that the month might change over the course of the week, we use the month of the first time-step only. Additionally, we extracted summary statistics from the following sensor time-series: raw acceleration, HR, HRV, Aceleration-derived Euclidean Norm Minus One, and Acceleration-derived Metabolic Equivalents of Task. Then, for every time-series we extracted the following variables which cover a diverse set of attributes of their distributions: mean, minimum, maximum, standard deviation, percentiles (25%, 50%, 75%), and the slope of a linear regression fit. The remainder of the variables (anthropometrics and RHR) are used as a single measurement.

In total, we derived a comprehensive set of 68 features using the Python libraries Pandas (Wes McKinney, 2010) and Numpy (Harris et al., 2020). A detailed view of the variables is provided in Table 6.1.

6.2.3 Deep learning models

We developed deep neural network models that are able to capture non-linear relationships between the input data and the respective outcomes. Considering the high-sampling rate of the sensors (1 sample/minute) after aligning HR and Acceleration modalities, it is impossible to learn patterns with such long dependencies (a week of sensor data includes more than 10,000 timesteps). Even the most well-tuned recurrent neural networks cannot cope with such sequences and given the size of the training set (7,545 samples), the best option was to extract statistical features from the sensors and represent every participant-week as a row in a feature vector (see Fig. 6.1). This feature vector was fed to fully connected neural network layers which were trained with backpropagation.

Data preparation. For Task 1 (see Figure 6.3), we match the sensor data with the participants who have eligible lab tests. Then we split into disjoint train and test sets, making sure that participants from *Fenland* I are allocated to the train set, while those from *Fenland* II are allocated to the test set (see Fig. 6.2). This would allow to re-use the trained model from Task 1, with different sensor data from *Fenland* II participants. Intuitively, we train a model on the big population, and we evaluate it with two snapshots of another longitudinal population over time (Task 1 & 3). After splitting, we normalize the training data by applying standard scaling (removing the mean

Table 6.1: **Description of the features/variables used in our analysis as inputs to the models.** The features with asterisks(*) are time-series and therefore we have extracted the following statistical variables: *mean, minimum, maximum, standard deviation, percentiles (25%, 50%, 75%), and the slope of a linear regression fit*. The final set of features is 68.

Features/Variables	Description
Sensors	
Acceleration*	Acceleration measured in mg
Heart rate (HR)*	Mean HR resampled in 15sec intervals, measured in BPM
Heart Rate Variability (HRV)*	HRV calculated by differencing the second-shortest and the second-longest inter-beat interval (as seen in (Faurholt-Jepsen et al., 2017)), measured in ms
Acceleration-derived Euclidean Norm Minus One (ENMO)*	ENMO-like variable, measured in (Acceleration/0.0060321) + 0.057) (as seen in (White et al., 2016))
Acceleration-derived Metabolic Equivalents of Task (METs)*	
Sedentary*	If Accelerometer <1 and daily count average
Moderate to Vigorous*	If Accelerometer >= 1 and daily count average
Vigorous*	If Accelerometer >= 4.15 and daily count average
Anthropometrics	
Age	Age, measured in years
Sex	Sex is binary (female/male)
Weight	Weight, measured in kilograms
Height	Height, measured in meters.centimeters
Body Mass Index (BMI)	BMI is calculated by Weight/(Height^2), measured in kg/ m2
Resting Heart Rate	
Wearable-derived RHR	RHR is calculated by averaging the 4th, 5th, and 6th minute of the baseline visit and adding to that the Sleeping Heart Rate that has been inferred by the wearable device. (Gonzales et al., 2020)
Seasonality	
Month of year	The month number is used along with a coordinate encoding that allows the models to make sense of their cyclical sequence.

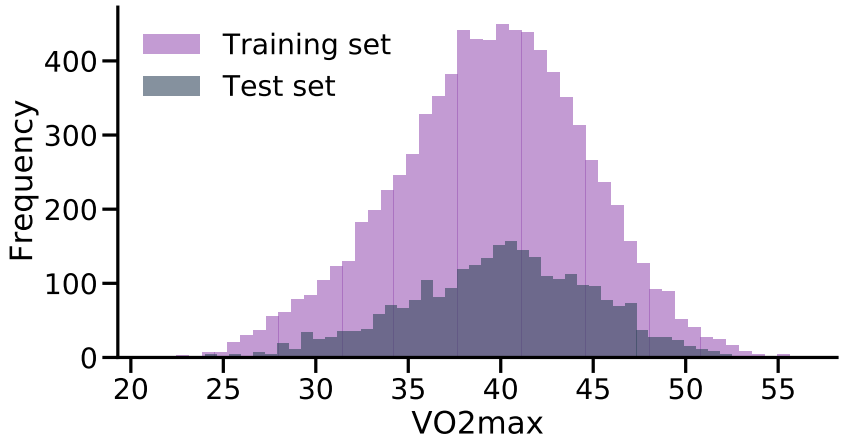


Figure 6.2: **Distribution of VO_2max in the training and test sets in *Fenland* I cohort.** Both sets display similar ranges of values, making sure that inferences based on the test set are robust. This plot refers to Task's 1 train and test sets.

and scaling to unit variance) and then denoise it by applying Principal Components Analysis (PCA), retaining the components that explain 99.99% of the variance. In practice, the original 68 features are reduced to 48. We save the *fitted* PCA projection and scaler and we apply them individually to the test-set, to avoid information leakage across the sets. The same projection and scaler are applied to all downstream models (Task 2 and 3) to leverage the knowledge of the big cohort (*Fenland* I).

Model architecture and training. The main neural network (used in Task 1) receives a 2D vector of [users, features] and predicts a real value. For this work, we assume N users and F features of an input vector $\mathbf{X} = (\mathbf{x}_1,...,\mathbf{x}_N) \in \mathbb{R}^{N \times F}$ and a target VO_2max $\mathbf{y} = (\mathbf{y}_1,...,\mathbf{y}_N) \in \mathbb{R}^N$. The network consists of two densely-connected feed forward layers with 128 units each. As we reviewed in Chapter 2, a dense layer works as follows: $output = activation\ (input \cdot kernel + bias)$, where activation is the element-wise activation function (the exponential linear unit in our case (Clevert et al., 2015)), kernel is a learned weights matrix with a Glorot uniform initialization (Glorot and Bengio, 2010), and bias is a learned bias vector. Each layer is followed by a *batch normalization* (Ioffe and Szegedy, 2015) operation, which maintains the mean output close to 0 and the output standard deviation close to 1. Also, dropout of 0.3 probability is applied to every layer, which randomly sets input units to 0 and helps prevent overfitting (Srivastava et al., 2014). Last, the final layer is a single-unit dense layer and the network is trained with the Adam optimizer (Kingma and Ba, 2014) to minimize the Mean Squared Error (MSE) loss, which is appropriate for continuous outcomes. We use a random 10% subset of the train-set as a validation set. To combat overfitting, we train for 300 epochs with a batch size of 32 and we perform early stopping when the validation loss stops improving after 15 epochs and the learning rate is reduced

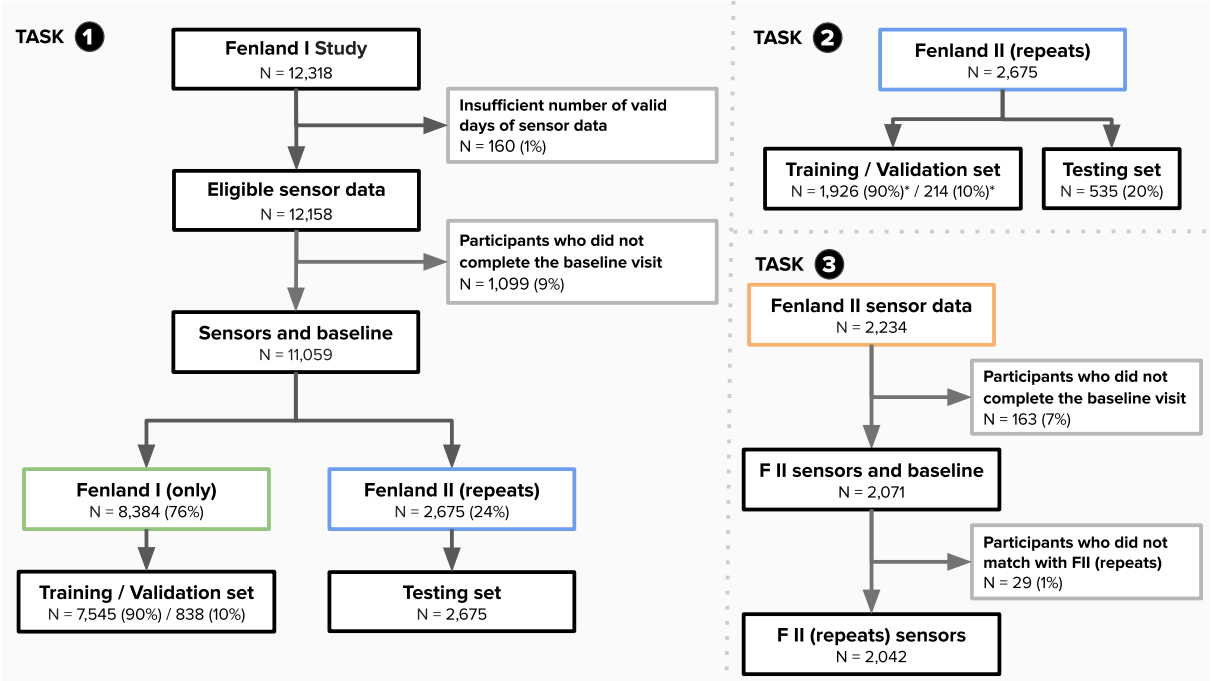


Figure 6.3: **Flowchart of the analytical sample and the training/testing splits across the three tasks.** The first task trains a model to predict fitness using the large cohort (*Fenland* I). The second task is using the smaller cohort of repeats in *Fenland* I (called *Fenland* II) and trains further models to predict fitness now and in the future (and their delta). The third task evaluates the original model trained in Task 1 by feeding new sensor data (*Fenland* II sensors and anthropometrics) to assess the adaptability of the model to pick up change. (*Training set is 90% of the 80% remaining dataset after splitting to testing set. Validation set is 10% of the training set)

by 0.1 every 5 epochs. All hyperparameters (# layers, # units, dropout rates, batch size, activations, and early stopping) were found after tuning on the validation set. We provide more details about the models in the Appendix A.

Model differences across tasks. Task 1 trains the main neural network of our study (see previous subsection). Task 2 re-trains an identical model to predict VO_2max in the future (and the delta present-future). However, when we re-frame this problem as a classification task (see Figure 6.5), we use significantly fewer participants when we focus on the tails of the change distribution. Therefore, to combat overfitting, we train a smaller network with only one Dense layer of 128 units and a sigmoid output unit, which is appropriate for binary problems. Instead of optimizing the MSE, we now minimize the binary cross-entropy. In all other cases — such as in Task 3 or when visualizing the latent space— we do not train new models; the model which was trained in Task 1 is used in inference mode (prediction).

6.2.4 Evaluation

To evaluate the performance of the deep learning models which predict continuous values, we computed the root mean squared error (RMSE) $= \sqrt{\frac{1}{|N_{test}|}\sum_{y\in\mathcal{D}_{test}}\sum_{t=1}^{N}(y_t-\hat{y}_t)^2}$, the coefficient of determination $(R^2) = 1 - \frac{\sum_{t=1}^{N}(y_t-\hat{y}_t)^2}{\sum_{t=1}^{N}(y_t-\bar{y}_t)^2}$, and the Pearson correlation coefficient. Here y and $\hat{y}$ are the measured and predicted VO_2max and $\bar{y}$ is the mean. For the binary models, we used the Area under the Receiver Operator Characteristic (AUROC or AUC) which evaluates the probability of a randomly selected positive sample to be ranked higher than a randomly selected negative sample.

6.2.5 Visualizing the latent space

The activations of the trained model allow us to understand the inner workings of the network and explore its latent space. We first pass the test-set of Task 1 through the trained model and retrieve the activations of the penultimate layer (Yosinski et al., 2015). This is a 2D vector of [2675, 128] size, considering that the layer size is 128 and the participants of the test-set are 2675. Intuitively, every participant corresponds to an 128-dimensional point. In order to visualize this embedding, we apply tSNE (Van der Maaten and Hinton, 2008), an algorithm for dimensionality reduction. For its optimization, we use a perplexity of 50, as it was suggested recently as an effective methodology (Wattenberg et al., 2016).

6.2.6 Statistical analyses

We performed a number of sensitivity analyses to investigate potential sources of bias in our results. Full results of these sensitivity analyses are shown in the main text and corresponding tables. In particular, we use bootstraping with replacement (1000 samples) to calculate 95% confidence intervals when we report the performance of the models in the hold-out sets (Carpenter and Bithell, 2000). Wherever we report p-values, we use the recently proposed strict threshold of $p < 0.005$ (Benjamin et al., 2018).

6.3 Results

Baseline measurements were collected from 12, 435 healthy adults from the *Fenland* study in the United Kingdom (Lindsay et al., 2019), where all required data for the present analysis was available in 11, 059 participants (*Fenland* I, baseline timepoint referred to as "current" in our evaluation). A subset of 2, 675 participants were assessed again after a median (interquantile range) of 7 (5-8) years (*Fenland* II, referred to as "future" in our evaluations). Descriptive characteristics of the two analysis samples

Table 6.2: **Characteristics for the study analytical sample: The *Fenland* I and II studies.** Data is in mean (std). Values with asterisk(*) indicate that this variable comes from *Fenland* II sensor data which is a smaller cohort (N=2071) due to data filtering (see Figure 6.3–Panel 3). The values in FII (future) cohort correspond to the second assessment (7 years later). In Task 1, the training set is FI (present) and the testing set is FII (present) so as to make sure that they come from similar distributions.

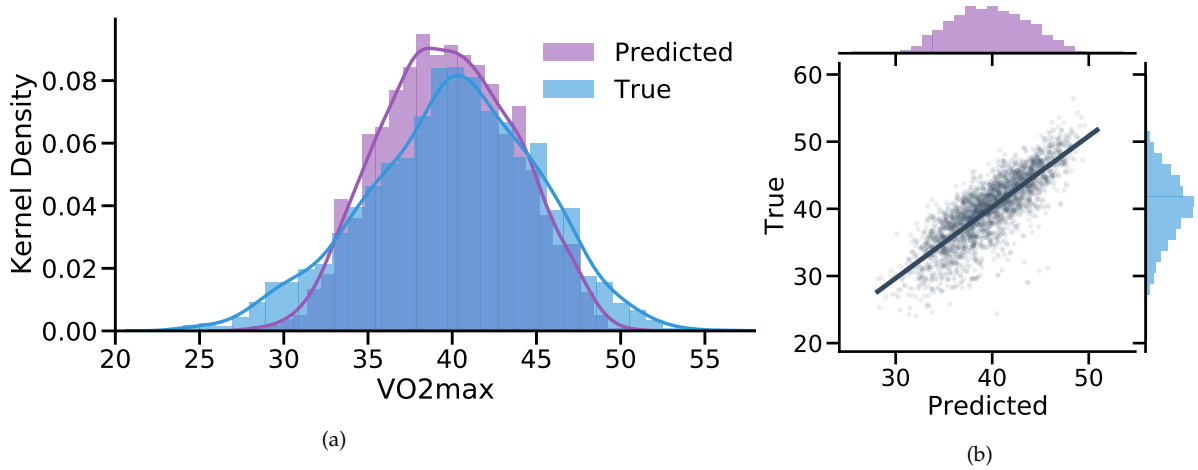
	***Fenland* $I_{present}$**				***Fenland* II_{future}**				***Fenland* $II_{present}$**			
	Men (n= 5229)		Women (n= 5830)		Men (n=1303)		Women (n=1372)		Men (n=1303)		Women (n=1372)	
	mean	*std*	*mean*	*std*	*mean*	*std*	*mean*	*std*	*mean*	*std*	*mean*	*std*
Demographics												
Age (years)	47.70	7.57	47.66	7.36	54.11	7.08	54.76	6.81	47.19	7.18	47.83	6.96
Anthropometrics												
Height (m)	1.78	0.07	1.64	0.06	1.77	0.06	1.64	0.06	1.77	0.06	1.64	0.06
Body mass (kg)	85.85	13.83	70.54	13.92	85.31	13.59	69.58	13.77	84.85	13.14	69.04	13.26
BMI (kg/m^2)	27.16	3.97	26.17	4.97	27.03	4.01	25.85	4.94	27.00	4.03	25.84	4.98
Physical activity												
MVPA (min/day)	35.87	22.35	34.40	22.59	34.92*	22.18*	35.35*	23.26*	34.41	22.23	32.81	21.45
VPA (min/day)	3.27	8.57	3.31	15.67	3.57*	8.78*	3.30*	7.52*	3.38	9.30	3.86	27.80
Resting Heart Rate												
RHR (bpm)	61.48	8.68	64.46	8.28	59.63*	8.28*	62.21*	8.10*	61.06	8.44	63.81	8.20
Cardiorespiratory fitness												
VO_2max (ml $O_2/min/kg$)	41.95	4.61	37.44	4.73	42.32	4.68	37.93	4.72	42.21	4.42	37.84	4.69

are presented in Table 6.2. Mean and standard deviations for each characteristic are presented in this table. An overview of the study design and the three experimental tasks is provided in Figure 6.1.

6.3.1 Fine-grained fitness prediction from wearable sensors

We first developed and externally validated several non-exercise VO_2max estimation models as a regression task using features commonly measured by wearable devices (anthropometry, resting heart rate (RHR), physical activity (PA); see Table 6.3). Here our goal was to explore how conventional non-exercise approaches to VO_2max estimation could be enhanced by features from free-living PA data. We split participant data into independent training and test sets. The training set (n=8384, participants with baseline data only) was used for model development. The test set (n=2675, participants with baseline and followup data) was used to externally validate each model. Models using anthropometry or RHR alone had poor external validity, but validity improved when combined in the same model. The best performance (R^2 of 0.67) was attained using a deep neural network model combining wearable sensors, RHR, and anthropometric data (Figure 6.4).

Deep neural networks can learn feature representations that are suitable for clustering tasks, such as population stratification by implicit health status, but are difficult to reveal using linear dimension-reduction techniques (Gaspar and Breen, 2019). We used t-distributed stochastic neighbor embedding (tSNE), a nonlinear dimension-reduction technique, to visualise deep-learned feature representations from our model and their



(a) (b)

Figure 6.4: **Fine-grained fitness prediction**. Comparing the predicted and true VO_2max coming from the best performing comprehensive model (*Sensors + RHR + Anthro.*) trained with *Fenland* I. **(a)** Distribution of predicted and true VO_2max. The plot combines a kernel density estimate and histogram. **(b)** Correlation of predicted and true VO_2max ($r = 0.82$, $p < 0.005$, see Table 6.3). The gray line denotes a linear regression fit. Transparency has been applied to the datapoints to combat crowding.

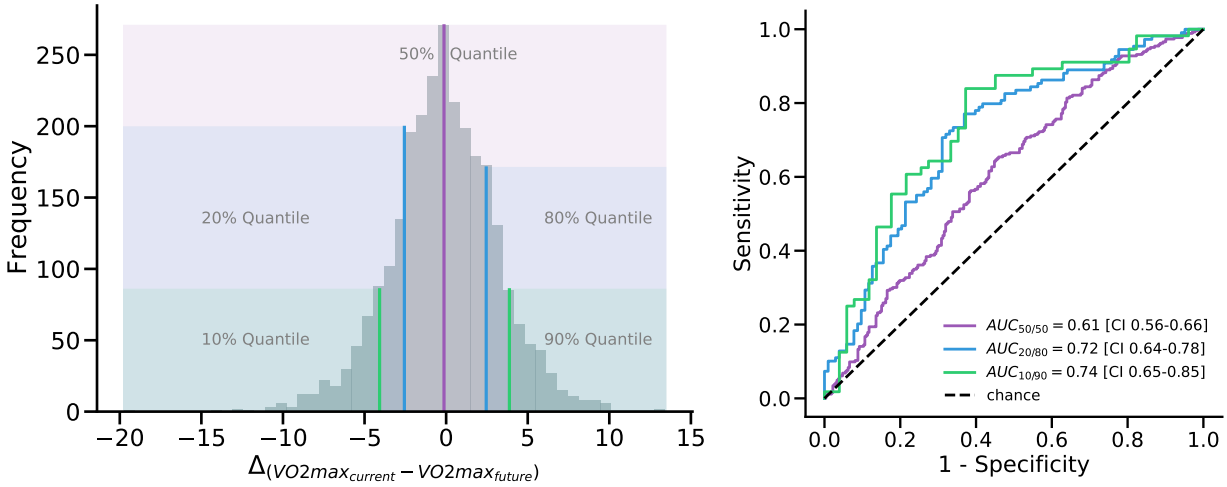
Table 6.3: **Evaluation of predicting fine-grained VO_2max with the *Fenland* I cohort**. Comparison between traditional antrhopometrics, common biomarkers (RHR), and passively collected data over a week (wearable sensors). Best performance in bold.

Data modality	Evaluation Metrics [95% CI]			N (train+val / test set)
	R^2	Corr	RMSE	
Anthropometrics				
Age/Sex/Weight/BMI/Height	0.362 [0.332-0.391]	0.604 [0.579-0.627]	4.043 [3.924-4.172]	
Resting Heart Rate				
RHR (Sensor-derived)	0.374 [0.344-0.403]	0.615 [0.589-0.639]	4.007 [3.891-4.117]	
Anthropometrics + RHR				11059
Age/Sex/Weight/BMI/Height/RHR	0.616 [0.588-0.641]	0.785 [0.767-0.802]	3.138 [3.031-3.237]	(8384/2675)
Wearable Sensors + RHR + Anthro.				
Acceleration/HR/HRV/MVPA	**0.671 [0.649-0.692]**	**0.822 [0.808-0.835]**	**2.903 [2.801-3.003]**	
Age/Sex/Weight/BMI/Height/RHR				

relationship to participant VO_2max and HR levels (Figure 6.6). Clustering and coloring by VO_2max and HR levels were shown to be inversely related and more apparent in the learned latent representation space, compared to the original observation space. For example, participants with higher VO_2max were clustered similarly to those with lower HR levels, and vice versa.

6.3.2 Predicting magnitude/direction of fitness change in the future

The second group of tasks evaluated our model on the subset of participants who returned for *Fenland* II approximately 7 years later (referred to as *future* in our evaluations). For these experiments we carried out three evaluations. Following the process



(a) (b)

Figure 6.5: **Evaluation in predicting the magnitude and direction of the VO_2max change between the present and the future. (a)** Distribution of the Δ of VO_2max in the present and the future. The shaded areas represent different binary bins that are used as outcomes, increasingly focusing on the extremes of this distribution. **(b)** ROC AUC performance in predicting the three Δ outcomes as shown on the left hand side. Brackets represent 95% CIs.

described earlier, we re-trained a model to predict future VO_2max using only information from the present as input (Table 6.4). This model yielded a slightly lower accuracy than *Fenland* I, achieving a R^2 of 0.49 and a correlation of 0.72. This lower performance is expected since the model has no indication of the behavior of the individuals 7 years later. We also trained a model to directly predict the difference (or delta) of current-future VO_2max, which reached a correlation of 0.23.

Further, motivated by the moderate predictability of the fine-grained delta of VO_2max, we formulated this problem as a classification task. A visual representation of this task can be found in Figure 6.5a. By inspecting the distribution of the difference (delta) of current-future VO_2max on the training set, we split it to 2 halves (50% quantiles) and set these as prediction outcomes. The purpose of this task is to assess the *direction* of individual change of fitness. We report an area under the curve (AUC) of 0.61 in predicting the direction of change ($N = 2,675$). We also focused on the tails of the change distribution which indicates participants who had substantial and dramatic change in fitness over the period of time between *Fenland* I and *Fenland* II (approximately 7 years). In this case, the distribution was split into 80%/20% (substantial) and 90%/10% (dramatic) quantiles. The results from these experiments show that the models can distinguish between substantial fitness change with an AUC of 0.72 ($N = 1,068$) and between dramatic fitness change with an AUC of 0.74 ($N = 535$). All AUC curves can be found in Figure 6.5b.

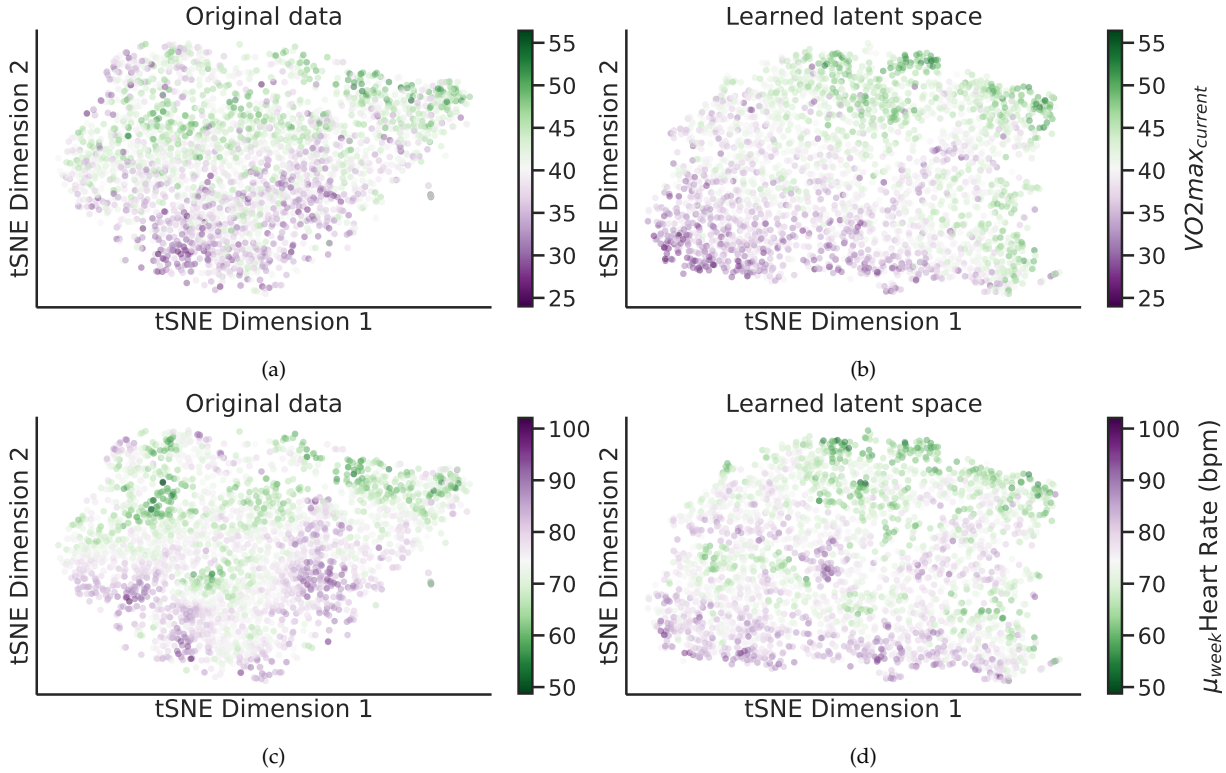


Figure 6.6: **t-distributed stochastic neighbor embedding (tSNE) projection of the original feature vector (*Fenland* I testing set, *Sensors* + *RHR* + *Anthro*.) compared to the model's latent space after training. (a-b)** The original data presents some clusters but the outcome is not clearly linearly separable. The model activations capture the continuum of low-high VO_2max both locally and globally. **(c-d)** A similar assessment to VO_2max is presented by coloring with the mean HR of the week of each participant. In the learned space, participants with low HR (high fitness) are placed in the same clusters as in VO_2max, unlike the original space. In all plots a 50% transparency has been applied to combat crowding and the colorbar is centered on the median value to illustrate extreme cases. The VO_2max label is used only for color-coding purposes (the projection is label-agnostic). Each participant is a dot.

Table 6.4: **Evaluation of predicting fine-grained VO_2max in the present and the future with the *Fenland* II repeats cohort using covariates of *Fenland* I**. Neural network results. (*the Delta outcome is in a different unit and hence a direct comparison with raw VO_2max results might not apply)

Outcomes	Evaluation Metrics [95% CI]			N (train+val / test set)
	R^2	Corr	RMSE	
Wearable Sensors + RHR + Anthro.				
Current VO_2max	0.652 [0.606-0.695]	0.815 [0.783-0.846]	2.959 [2.742-3.201]	2675 (2140/535)
Future VO_2max	0.499 [0.431-0.55]	0.721 [0.67-0.759]	3.673 [3.421-3.916]	
Delta (Current - Future)*	0.081 [0.02-0.078]	0.233 [0.159-0.307]	3.175 [2.923-3.41]	

6.3.3 Enabling adaptive cardio-respiratory fitness inferences

For the final task, we assessed whether the trained models can pick up change using new sensor data from *Fenland* II, considering that obtaining new wearable data is relatively easy since these devices are becoming increasingly pervasive. The intuition behind this task is to evaluate the generalizability of the models over time. We first matched the populations that provided sensor data for both cohorts ($N = 2,042$) and applied the trained model from Task 1 in order to produce VO_2max inferences. We then compared the predictions with the respective ground truth (current and future VO_2max). The true and predictive distributions are shown in Figures 6.7c and 6.7d. Through this procedure, we found that the model achieves an $r = 0.84$ for VO_2max future prediction and an $r = 0.82$ for VO_2max current prediction (validating our Task 1 results). In other words, if we have access to wearable sensor data and other information from the future time, we can reuse the already trained model from *Fenland* I to accurately infer fitness with minimal loss of accuracy over time, even though this is new sensor data from a completely separate (future) week.

Last, we calculated the delta of the predictions and compared it to the actual delta of fitness over the years. This task showed that the models tend to focus mostly on positive change and under-predict when participants' fitness deteriorates over the years (Figures 6.7a, 6.7b). The overall correlation between the delta of the predictions with the ground truth is significant (r=0.57, $p<0.005$).

6.3.4 Contextualising the results

Cardiorespiratory fitness declines with age independently of changes to body composition, and low cardiorespiratory fitness is associated with poor health outcomes (Lynch et al., 1996; Lakka et al., 1994; Myers et al., 2002; Ekelund et al., 1988; Schmid and Leitzmann, 2015; Schuch et al., 2016; Blair et al., 1989; Laukkanen et al., 2004; Mandsager

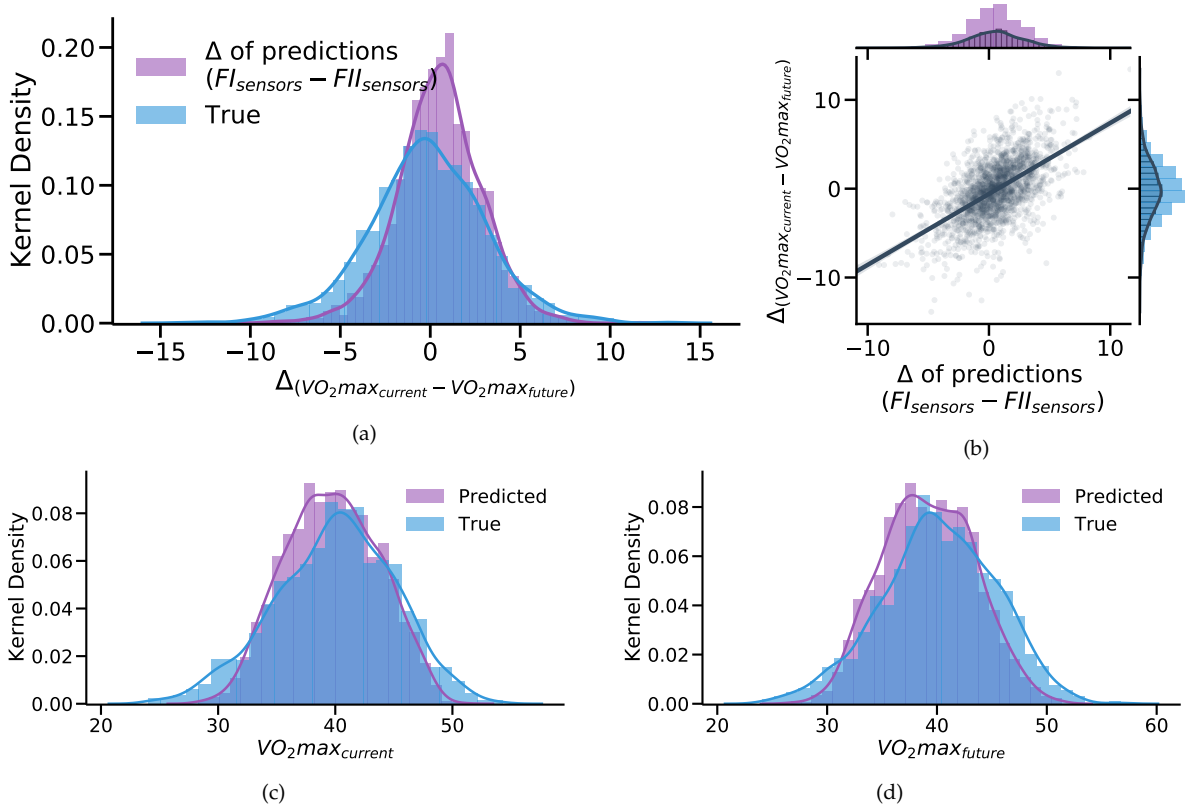


Figure 6.7: **Assessing the robustness of the model to pick up change using new sensor data from *Fenland* II repeats**. By matching the populations who provided sensor data for both cohorts (N=2,042) we passed them through our trained model from Task 1 to predict VO_2max. **(a-b)** We then calculated the difference (Δ) of the predictions juxtaposed with the true difference of fitness over the years. Distribution of Δ of predicted and true VO_2max. Correlation of Δ of predicted and true VO_2max ($r = 0.57$, $p < 0.005$). The gray line denotes a linear regression fit. Transparency has been applied to the datapoints to combat crowding. **(c-d)** Comparison of predicted and true VO_2max using FI and FII covariates (sensors, RHR, anthro.), respectively. The distribution plots combine a kernel density estimate and histogram with bin size determined automatically with a reference rule.

et al., 2018). As such, having the capacity to predict whether CRF would decline in excess of natural aging could be valuable to clinicians when tailoring therapeutic interventions. Here we have developed a deep learning framework for predicting CRF and changes in CRF over time. Our framework estimates VO_2max by combining learned features from heart rate and accelerometer free-living data extracted from wearable sensors with anthropometric measures. To evaluate our framework's performance, VO_2max estimates were compared with VO_2max values derived from a submaximal exercise test (Gonzales et al., 2020). Free-living and exercise test data were collected at a baseline investigation in 11,059 participants (*Fenland* I). A subset of those participants (n=2,675) completed another exercise test at a follow-up investigation approximately seven years later (*Fenland* II). This study design allowed us to address three questions: 1) Do baseline estimates of VO_2max from the deep learning framework agree with VO_2max values measured from exercise testing at baseline?, 2) Can the framework learn features from heart rate and accelerometer free-living data collected at baseline that predict VO_2max measured at follow-up?, and 3) Can the framework be used to predict the magnitude of change in VO_2max from baseline to follow-up?

In the VO_2max estimation tasks, our model demonstrated strong agreement with VO_2max measured from the submaximal exercise test at baseline (Pearson's correlation coefficient (PCC): 0.82) as well as for the longitudinal, follow-up visit (PCC: 0.72). We were also able to distinguish between substantial and dramatic changes in CRF (AUCs 0.72 and 0.74, respectively). Finally, we further evaluated the initial model on new input data by feeding *Fenland* II free-living data along with updated heart rate and anthropometrics to the model, showing that it is able to adapt and monitor change over time. We evaluated the inference capabilities of the model in the difference (delta) between the current (*Fenland* I) and future (*Fenland* II) VO_2max for those participants who came back approximately 7 years later. For this last task, the model produced outcomes that translated to a 0.57 correlation between the delta of predicted and delta of true VO_2max.

The application of our work to other cohort and longitudinal studies is of particular importance because serial measurement of cardiorespiratory fitness has significant prognostic value in clinical practice. Small increases in fitness are associated with reduced cardiovascular disease mortality risk (Blair et al., 1995) and better clinical outcomes in patients with heart failure (Swank et al., 2012) and type 2 diabetes (Jakicic et al., 2013). Nevertheless, routine measurement of fitness in clinical practice is rare due to the costs and risks of exercise testing. Non-exercise based regression models can be used to estimate changes in fitness in lieu of serial exercise testing. It is unclear, however, the extent to which changes in fitness detected with such models reflect true changes in exercise capacity. Here, we relied on the relationship between CRF and

heart rate responses to different levels of physical activity at submaximal, real-life conditions captured through wearable sensors. Using deep learning techniques, we have developed a non-exercise based fitness estimation approach that can be used not only to accurately infer current VO_2max, but also can do so when applied to a future cohort, where the model did not require any retraining, just influx of new data. Further, we show that the model can also be used to infer the changes in CRF that occurred during the approximately seven year time span between *Fenland* I and II.

Our proposed deep learning approach outperforms traditional non-exercise models, which are the state-of-the art in the field and rely on simple variables inputted to a linear model. Importantly, our model is able to take week-level information from each participant and combine it with various anthropometrics and biomarkers such as the RHR, providing a truly personalized approach for CRF inference generation. The approach we present here outperforms traditional non-exercise models, which are considered state-of-the-art methods for longitudinal monitoring and highlights the potential of wearable sensing technologies for digital health monitoring.

This study has several limitations worthy of recognition. First, the validity of the deep learning framework was assessed by comparing estimated VO_2max values with those derived from a submaximal exercise test. Ideally, one would use VO_2max values directly measured during a maximal exercise test to establish the ground truth for cardiorespiratory fitness comparisons. Maximal exercise tests, however, are problematic when used in large population based studies because they may be unsafe for some participants and, consequently, induce selection bias. The submaximal exercise test used in the *Fenland* Study was well-tolerated by study participants and demonstrated acceptable validity against direct VO_2max measurements (Gonzales et al., 2020). We are therefore confident that VO_2max values estimated from the deep learning framework reflect true cardiorespiratory fitness levels.

6.4 Discussion

Although the use of wearable devices continues to grow, most of the derived variables in commercial wearable devices lack rigorous scientific validation and as such, their use in health-related inferences has been questioned (Henriksen et al., 2018; Passler et al., 2019; Shcherbina et al., 2017; Boudreaux et al., 2018; Perez-Pozuelo et al.). Specifically, VO_2max estimations using commercial devices are particularly non-transparent and at times unreliable (Shcherbina et al., 2017; Esco et al., 2011). Although certain commercial devices have shown stronger results than others, many tend to rely on detailed activity intensity measurements paired with speed monitoring through GPS and require users to reach heart rates that are close to their maximum capabilities, limiting the application

to self-selecting, fitter individuals (Cooper and Shafer, 2019; Lucio et al., 2018). Despite some promising studies which attempt to infer VO_2max from data collected during free-living conditions, these mostly stem from small-scale cohorts with less than 50 participants and use contextual data from treadmill activity, which again limits their application in real-world contexts (Altini et al., 2016). In this work, we use data from the largest study of it's kind, by over two orders of magnitude, and use purely free-living data to predict VO_2max, with no requirement for context-awareness.

6.5 Conclusion

In this chapter, which concludes the empirical works of this thesis, we developed deep learning models utilising wearable data and other biomarkers to predict the gold standard of fitness (VO_2max) and achieved strong performance compared to other traditional approaches. Cardio-respiratory fitness is a well-established predictor of metabolic disease and mortality and our premise is that modern wearables capture non-standardised dynamic data which could improve fitness prediction. Our findings on a population of 11, 059 participants showed that the combination of all modalities reached an $r = 0.82$, when compared to the ground truth in a holdout sample. Additionally, we show the adaptability and applicability of this approach for detecting fitness change over time in a longitudinal subsample ($n = 2, 675$) who repeated measurements after seven years. Last, the latent representations that arise from this model pave the way for fitness-aware monitoring and interventions at scale. It is often said that *"If you cannot measure it, you cannot improve it"*. Cardio-respiratory fitness is such an important health marker, but until now we did not have the means to measure it at scale. Our findings could have significant implications for population health policies, finally moving beyond weak health proxies such as the BMI.

This chapter built on ideas of the previous chapters, such as the promising predictability of fitness in Chapter 5, the task-inspired feature extraction in Chapter 3, and the latent patterns seen in the intermediate representations in Chapters 4 & 5. While this chapter featured the most medically-relevant application of the four, we believe that ideas from the rest of the chapters can be applied here as well. For example, we did not see significant improvements in predicting fine-grained VO_2max when using a self-supervised formulation similar to that of Chapter 5, which can be attributed to the difference in devices since here we use a uniaxial chest accelerometer while in Chapter 5 the input data was a wrist-worn triaxial accelerometer. However, we still believe that other self-supervised objectives such as in SimCLR (Chen et al., 2020; Tang et al., 2020) should be beneficial in learning robust representations of large-scale sensor data.

Chapter 7

Conclusion

Sometimes it seems as though each new step towards AI, rather than producing something which everyone agrees is real intelligence, merely reveals what real intelligence is not.
–Douglas R. Hofstadter

In this thesis, we presented four original pieces of work drawing on some fundamental research problems in machine learning for mobile health: finding better data representations through neural networks and validating the impact of sensor data when compared to other traditional sources. Our premise has been that new training paradigms such as multi-tasking, self-supervision, and multimodal machine learning should create more robust predictive models, which in turn can be applied to tasks in mental and physical health. In this Chapter, we briefly summarize our key contributions and suggest directions for future research.

7.1 Summary of contributions

7.1.1 Multimodal machine learning for mood prediction

In Chapter 3, we presented a training pipeline for population-scale mobile sensor data towards more accurate mood clustering and prediction. The main motivation behind this study was that experience sampling has been proposed as a mechanism to monitor mental health, but it requires users' attention and therefore this leads to considerable retention issues. The proposed training pipeline involved two steps: *first*, clustering historical mood trajectories in order to find groups of users with similar trajectories and *second*, classifying users into the found clusters. We found that the combination of these modalities achieves the best classification performance, and that passive sensing yields a +5% boost in performance (75% AUC). These findings might have implications for digital phenotyping applications that can benefit from the correct modelling of large-scale passive sensing data alongside extra user metadata.

7.1.2 Sequence multi-task learning for mood forecasting

In Chapter 4, we presented an encoder-decoder model which exploits the bimodality of mood with multi-task learning, enabling more accurate multi-step mood forecasting. Our results showed that multi-tasking learns both dimensions of mood simultaneously, which is more accurate than individual models or baselines. Our results showed that 3 weeks is the best window of mood reporting, validating previous research on depression prediction. Also, our models outperformed regressors and other baselines, while extra analysis showed that mood variability, personality traits, and that the day of the week play a key role in the performance of the models. Last, we inspected the learned representations and observed that different neurons learn different non-linear sequential patterns, which helps us understand the complex trajectories of the evolution of mood.

7.1.3 Self-supervised transfer learning for wearable data

In Chapter 5, we developed a self-supervised model which exploits the multimodal data of modern wearables to learn meaningful representations which generalize to several outcomes with transfer learning. The model maps activity data to heart rate responses and can be used as a feature extractor for wearable data. For pre-training, we introduced a joint loss function that accounts for the long-tails of HR data, while downstream, we aggregated the window-level features to user-level ones and showcased the value captured by the learned embeddings through strong performance at inferring physiologically meaningful variables, outperforming autoencoders and common biomarkers. For example, our models achieved an AUC of 0.70 for BMI prediction and an AUC of 0.80 for Physical Activity Energy Expenditure.

7.1.4 Longitudinal fitness prediction with wearables

In Chapter 6, we developed deep learning models utilising wearable data and common biomarkers to predict the gold standard of fitness (VO_2max) and achieved strong performance compared to traditional approaches. Cardio-respiratory fitness is a well-established predictor of metabolic disease and mortality and our premise is that modern wearables capture non-standardised dynamic data which could improve fitness prediction. Our findings on a population of 11, 059 participants showed that the combination of all modalities reached an $r = 0.82$, when compared to the ground truth in a holdout sample. Additionally, we show the adaptability and applicability of this approach for detecting fitness change over time in a longitudinal subsample ($n = 2, 675$) who repeated measurements after seven years. Last, the latent representations

that arise from this model pave the way for fitness-aware monitoring and interventions at scale.

7.2 Implications and limitations

The work presented in this thesis has potential implications for various communities and stakeholders. Researchers could use methods, ideas, and developed models to produce inferences and predictions for their own data and study different populations. Engineers could create ML/software products targeted to mobile and wearable devices which understand the context of the users by anticipating mental health instabilities and the link of physical activity to metabolic health. On the other hand, medical practitioners could use the outputs of such software products to better understand their patients' daily lives away from hospital settings; continuous, passive person-generated data can complement episodic data generated during routine clinical practice. Last, policymakers could use our findings to advocate for new population health initiatives; for example, if a cheap wearable device can offer better proxies for one's overall health and mortality than demographics or aggregated self-reported metrics (e.g. BMI), they could support nationwide initiatives similar to "One Laptop per Child".

All studies have limitations. Our work is potentially affected by the nature of the data at hand. Observational studies provide larger samples which more closely approximate the general population but, at the same time, the researchers cannot control any interventions or exposures. For example, considering that our mental health dataset was collected through a widely distributed mobile app, we can assume some selection bias from people who tend to follow science news and live in Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies (Henrich et al., 2010). However, it is still today one the most comprehensive global datasets to study the interplay of mobile sensing and mental health.

Further, our physical activity dataset was focused on a regional population in England, but due to being a prospective population-based cohort study, it included more control in order to ensure that the sample is nationally representative. We also acknowledge the fact that this dataset was originally collected to study metabolic disease in people born from 1950 to 1975, therefore our results might not generalize to very young populations. On the other hand, predicting poor overall health through fitness biomarkers is particularly more significant in older populations due to increased mortality risk.

Another limitation refers to *missing modalities*, namely over-relying on data inputs which might be easy to collect but do not explain the underlying research question. For instance, our mental health dataset includes movement sensors, background audio,

personality surveys and other mobile phone metadata. External factors such as the weather or menstruation cycles for women could also influence mood (Golub and Harrington, 1981), but the collection of weather data is challenging on a globally distributed user-base and menstruation could be regarded as very private information to self-report. Likewise, fitness is influenced by both physical activity and nutrition, but there is considerable bias in self-reported meal intake (Schoeller, 1995) (besides being impractical to log).

Another important point worth mentioning is the degree in which the proposed models can generalise to other domains. We discussed in previous chapters that the models developed for mental health outcomes could be also used for physical activity outcomes and vice versa. Given that the main focus of this thesis is on modeling sequential data, we believe that the models could potentially be applied to other domains where parallel time-series are prevalent — such as energy or traffic — in order to learn rich cross-modal representations (as in Chapter 5). Besides, in the last year we have witnessed a remarkable consolidation in deep learning architectures, with most modalities being modeled with a variant of the Transfomer architecture (Dosovitskiy et al., 2020) — we expect this trend to accelerate in the future.

All things considered, there is no silver bullet for this kind of challenges, however, we believe that by carefully formulating the research questions and slicing the data in a meaningful way which respects the temporal/causal aspect thereof, can alleviate some of these problems. In our studies, we employ forecasting and user-based cross-validation, as well as we validate our models in future cohorts repeating the same protocols, as a means of reducing bias.

7.3 Future research directions

Paraphrasing the quote in the beginning of this chapter, every step towards more intelligent machines reveals limitations which mostly arise from the way that we formulate such prediction tasks. As such, a simplistic task may yield impressive accuracy, which however will not generalise to the real world. Future models should be able to exploit different modalities, limited ground truth, and discover hidden causal effects in person-generated observational data.

Nevertheless, we should keep in mind that these problems are *hard*. For instance, Moravec's paradox states that, contrary to traditional beliefs, reasoning (*"playing checkers"*) requires very little computation, whereas sensorimotor capabilities (*"the skills of a one-year-old when it comes to perception and mobility"*) require enormous resources (Moravec, 1998). Even though the original observation referred to robotics, we could extrapolate it to every field that requires intelligence from sensory inputs. In other

words, this observation helps us understand why AI has been first successful in strictly-defined reasoning tasks (e.g., chess, Go), over open-ended problems that involve perception. On top of that, mobile health -apart from the sensory component- includes another layer of ambiguity, where there might not be strong consensus in what constitutes an outcome (e.g., depression).

The above remarks should motivate more work with real data and tasks. As such, below are some potential future directions.

7.3.1 Multimodal health modeling: striking a fine balance

In most of the Chapters we used modalities such as the accelerometer, the electrocardiogram, or the microphone. However, as humans employ all their senses when navigating the word, we could expand to underutilized modalities, but as in most cases, we are limited by the existing datasets. For example, it has been recently suggested that the task of human activity recognition might have reached a plateau by focusing on accelerometers only, with the authors proposing a vision-based alternative (Tong et al., 2020). Admittedly, every new modality complicates the fine balance between privacy and accuracy, while motivating new policy frameworks for data governance (Perez-Pozuelo et al., 2021). Still, we need new models which can ingest structurally different data types (e.g., video and signals) in a principled way; large pre-trained models (Bommasani et al., 2021) and contrastive learning have a lot to offer here (Wang et al., 2021).

If for a moment we imagine that we have the perfect multimodal datasets, it is not yet straightforward how to optimally fuse different modalities. Recently, a new category was coined to describe models that are trained on "*broad data at scale and are adaptable to a wide range of downstream tasks*", the authors of the report call them *Foundation* models (Bommasani et al., 2021). In this ideal scenario, data from care providers, institutions (universities, non-profits, and governments), pharmaceuticals, wearables, and insights extracted from medical publications would be sourced. Then, individual modalities would be extracted including medical images, ultrasound videos, tabular electronic health records (EHRs) data, text from clinical notes, and time series such as ECGs from wearables. As these models are particularly adaptable through fine-tuning and prompting, they can be used in all sorts of useful downstream tasks such as question answering by both doctors and patients.

Nevertheless, current self-supervised models are developed for each modality independently, e.g., images (Chen et al., 2020), text (Devlin et al., 2019), and ECG (Kiyasseh et al., 2020). Therefore, we need methods that learn cross-modal patterns on different fusion levels (patient, population, and temporal). A promising recent direction

is surpisingly one which is modality-agnostic; the *Perceiver* is a Tranformer which maps input arrays[1] to output arrays through a small latent array (Jaegle et al., 2021). To illustrate this contribution, they employed a dataset with audio-video-label inputs and the objective was to learn a model that can accurately compress and reconstruct its inputs in the presence of a bottleneck. With traditional autoencoders like CNNs, it is not obvious how to fuse these three modalities since video is 3D, audio is 1D and class labels are 0D. The *Perceiver* team used padded inputs, serialized them into a 2D array, and queried the model using Fourier-based position encodings (Tancik et al., 2020). Other tricks involved masking the label 50% of the time (similar to BERT) and subsampling the decoding, both during training. The results showed consistently low peak signal-to-noise ratios for both audio and video which hints that the model learns a joint distribution across modalities.

These types of input-agnostic models have the potential to automate the manual process of building multi-modal pipelines, and in particular mobile health is well-poised to benefit from models that learn cross-modal patterns of tabular, sequential, and spatial data.

7.3.2 Representation learning: contrastive, generative, or both?

As we discussed back in Chapter 2, unsupervised models for the first time outperformed supervised ones (Lan et al., 2020), even in sensor tasks (Saeed et al., 2019; Tang et al., 2021). Generic pre-training methods like *SimCLR* (Chen et al., 2020) or *BYOL* (Grill et al., 2020) proposed a two-network setup which ingests different views of the same datapoint, with the loss minimizing the distance of the latent representations. This sounds surprisingly similar to another family of models: Generative Adversarial Networks (GANs), where the objective draws from game-theoretic principles and two networks contest with each other in a game to generate more realistic data. We expect to see more overlap between generative and contrastive training in the future. We point the interested readers to this survey on the similarities and differences of these two paradigms (Liu et al., 2020). Beyond generic training, we are also excited about timeseries-specific self-supervised models which take into account properties such as local smoothness (Tonekaboni et al., 2021), and other spatio-temporal invariances (Kiyasseh et al., 2020).

Starting from the latter, we acknowledge that ECG data is ubiquitous in healthcare settings and is increasingly common in personal devices such as the Apple Watch. *CLOCS* proposed a method which leverages temporal and spatial invariances of ECG leads based on the two key observations: adjacent ECG segments of shorter duration

[1] a byte array can be a flattened image or an entire ECG sequence and is generally large

will continue to share context, and recordings from different leads (at the same time) will reflect the same cardiac function, and thus share context (Kiyasseh et al., 2020). The new idea was to define a *positive pair* as representations of transformed instances that belong to the same patient. By doing so, the model implicitly personalizes the learned representations to each patient. Driven by this, they designed a new SimCLR-like objective that outperformed supervised and generic self-supervised methods (in terms of AUC) such as BYOL, most notably, with only 25% of labelled training data.

Yèche et al took the idea of inducing priors on contrastive losses a step forward (Yèche et al., 2021). They design an objective that preserves the time dependency of the representations of the time-series segments and outperforms unsupervised and supervised methods in predicting ICU decompensation, length of stay, and sepsis onset (on the MIMIC dataset). The versatility of this approach is twofold: when fully unsupervised, it is competitive to supervised models, and when used in a supervised manner, it outperforms contrastive methods. Another study independently arrived to a similar formulation (Tonekaboni et al., 2021) by ensuring that in the encoding space, the distribution of signals from within a neighborhood is distinguishable from the distribution of non-neighboring signals. Their models surpassed competitors such as the Triplet Loss and Contrastive Predictive Coding in predicting diverse outcomes ranging from atrial fibrillation to human activity recognition. They also showed better clusterability over other contrastive losses. Both studies highlighted the generality of such models which can be reused in multiple downstream tasks.

Another study proposed a self-supervised model with an adversarial subject identifier to minimize subject-specific content (Cheng et al., 2020). We expect these subject-focused objectives to become more common when training unsupervised models on large health datasets. For a comprehensive view of the field, we point the reader to our recent review (Spathis et al., 2022).

7.3.3 Transfer learning: the next frontier of intelligent health?

Andrew Ng has recently expressed some alarm that there is a a considerable gap between proof of concept models and actual in-situ use, due to different sensors, protocols, or data collection methods: *"In contrast, any human [..] can walk down the street to the other hospital and do just fine"* (Reader, 2021). Sequential transfer learning, as seen for example in Chapter 5 or in numerous recent works (Grill et al., 2020), is probably the first step to validate that the learned representations can generalize across many different tasks. We should now go the extra mile and validate that these models can perform equally well in changing environments. Some exciting new approaches towards this direction include disentangled and adversarial autoencoders

(Han et al., 2021), domain adaptation (Yang and Soatto, 2020), and meta learning (Li and Hospedales, 2020).

A potential fertile ground for improving transfer learning methods is through domain generalization (DG). DG deals with the problem of out-of-distribution generalization and has attracted increasing interest (Wang et al., 2021). In this setup, several different but related domains are given, and the goal is to learn a model that is invariant to the input domains and can generalize to unseen test domains. Conceptually different approaches have been proposed to solve this problem, with the most noteworthy being Invariant Risk Minimization (IRM) (Arjovsky et al., 2019), which aims to extend traditional domain-unaware training or Empirical Risk Minimization (ERM). IRM features a joint loss function: the ERM component which tries to minimize the average risk across all environments, and the IRM one optimizes the data representation such that all domains have the same downstream classifier. Another approach applies the meta-learning paradigm to the DG setting: MLDG simulates train/test domain shifts during training by synthesizing virtual testing domains within each mini-batch, with a meta-objective that assumes improvement in training domain performance will reflect on test domain performance (Li et al., 2018). Other ideas such as GroupDRO focused on minimizing the worst-case training loss of each domain (Sagawa et al., 2019), and CORAL suggested aligning the mean and covariance of latent distributions across domains (Sun and Saenko, 2016).

Even though all these methods produced superior results on their respective evaluation setups, recent benchmarks on a large array of datasets and methods criticized their effectiveness over simple baselines like the ERM (Gulrajani and Lopez-Paz, 2020). Later works challenged these findings, claiming that progress has actually been made over ERM, pointing to pre-training and augmentations (learned/generative or heuristic) as potential solutions (Wiles et al., 2021). However, the culprit behind these inconsistencies could be the lack of high-quality distribution shift datasets and benchmarks. While steps have been taken to introduce new multi-domain ML benchmarks (Koh et al., 2021), they heavily feature vision and language modalities, with limited support for timeseries (let alone medical or wearable timeseries). We are particularly excited about new methods and benchmarks in this area because it could unlock the full potential of generalization to unseen domains with versatile methods that learn the essence of data, regardless of populations, devices, and environments.

Bibliography

[1] Georges Aad, Tatevik Abajyan, B Abbott, J Abdallah, S Abdel Khalek, Ahmed Ali Abdelalim, R Aben, B Abi, M Abolins, OS AbouZeid, et al. Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc. *Physics Letters B*, 716(1):1–29, 2012.

[2] Alireza Abedin, S. Hamid Rezatofighi, Qinfeng Shi, and Damith C. Ranasinghe. Sparsesense: Human activity recognition from highly sparse sensor data-streams using set-based neural networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5780–5786. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/801. URL https://doi.org/10.24963/ijcai.2019/801.

[3] Fatih Abut, Mehmet Fatih Akay, and James George. Developing new vo2max prediction models from maximal, submaximal and questionnaire variables using support vector machines combined with feature selection. *Computers in biology and medicine*, 79:182–192, 2016.

[4] Karan Aggarwal, Shafiq Joty, Luis Fernandez-Luque, and Jaideep Srivastava. Adversarial unsupervised representation learning for activity time-series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 834–841, 2019.

[5] Stephen Aichele, Patrick Rabbitt, and Paolo Ghisletta. Think fast, feel fine, live long: A 29-year study of cognition, health, and survival in middle-aged and older adults. *Psychological science*, 27(4):518–529, 2016.

[6] Mohammad Abu Alsheikh, Ahmed Selim, Dusit Niyato, Linda Doyle, Shaowei Lin, and Hwee-Pink Tan. Deep activity recognition models with triaxial accelerometers. *arXiv preprint arXiv:1511.04664*, 2015.

[7] Tim Althoff, Jennifer L Hicks, Abby C King, Scott L Delp, Jure Leskovec, et al. Large-scale physical activity data reveal worldwide activity inequality. *Nature*, 547(7663):336, 2017.

[8] Marco Altini, Pierluigi Casale, Julien Penders, and Oliver Amft. Cardiorespiratory fitness estimation in free-living using wearable sensors. *Artificial intelligence in medicine*, 68:37–46, 2016.

[9] Wei Tech Ang, Pradeep K Khosla, and Cameron N Riviere. Nonlinear regression model of a low-g mems accelerometer. *IEEE Sensors Journal*, 7(1):81–88, 2007.

[10] Sinan Aral and Christos Nicolaides. Exercise contagion in a global social network. *Nature communications*, 8(1):1–8, 2017.

[11] Charles S Areni and Mitchell Burger. Memories of "bad" days are more biased than memories of "good" days: past saturdays vary, but past mondays are always blue. *Journal of Applied Social Psychology*, 38(6):1395–1415, 2008.

[12] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[13] Felix Assah, Jean Claude Mbanya, Ulf Ekelund, Nicholas Wareham, and Soren Brage. Patterns and correlates of objectively measured free-living physical activity in adults in rural and urban cameroon. *J Epidemiol Community Health*, 69 (7):700–707, 2015.

[14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[15] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[16] 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. Deepheart: Semi-supervised sequence learning for cardiovascular risk prediction. *arXiv preprint arXiv:1802.02511*, 2018.

[17] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5 (2):157–166, 1994.

[18] Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. Redefine statistical significance. *Nature human behaviour*, 2 (1):6–10, 2018.

[19] Brinnae Bent, Benjamin A Goldstein, Warren A Kibbe, and Jessilyn P Dunn. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ digital medicine*, 3(1):1–9, 2020.

[20] Steven N Blair, Harold W Kohl, Ralph S Paffenbarger, Debra G Clark, Kenneth H Cooper, and Larry W Gibbons. Physical fitness and all-cause mortality: a prospective study of healthy men and women. *Jama*, 262(17):2395–2401, 1989.

[21] Steven N Blair, Harold W Kohl, Carolyn E Barlow, Ralph S Paffenbarger, Larry W Gibbons, and Caroline A Macera. Changes in physical fitness and all-cause mortality: a prospective study of healthy and unhealthy men. *Jama*, 273(14): 1093–1098, 1995.

[22] Davis W Blalock and John V Guttag. Extract: Strong examples from weakly-labeled sensor data. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 799–804. IEEE, 2016.

[23] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Sandy Pentland. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 477–486. ACM, 2014.

[24] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[25] Alberto G Bonomi, Gill A Ten Hoor, Helma M De Morree, Guy Plasqui, and Francesco Sartor. Cardiorespiratory fitness estimation from heart rate and body movement in daily life. *Journal of Applied Physiology*, 128(3):493–500, 2020.

[26] Benjamin D Boudreaux, Edward P Hebert, Daniel B Hollander, Brian M Williams, Corinne L Cormier, Mildred R Naquin, Wynn W Gillan, Emily E Gusew, and Robert R Kraemer. Validity of wearable activity monitors during cycling and resistance exercise. *Medicine and science in sports and exercise*, 50(3):624–633, 2018.

[27] Søren Brage, Niels Brage, Paul W Franks, Ulf Ekelund, Man-Yu Wong, Lars Bo Andersen, Karsten Froberg, and Nicholas J Wareham. 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):343–351, 2004.

[28] Søren Brage, Ulf Ekelund, Niels Brage, Mark A Hennings, Karsten Froberg, Paul W Franks, and Nicholas J Wareham. Hierarchy of individual calibration levels for heart rate and accelerometry to measure physical activity. *Journal of Applied Physiology*, 103(2):682–692, 2007.

[29] Søren Brage, Kate Westgate, Paul W Franks, Oliver Stegle, Antony Wright, Ulf Ekelund, and Nicholas J Wareham. Estimation of free-living energy expenditure by heart rate and movement sensing: a doubly-labelled water study. *PloS one*, 10 (9):e0137206, 2015.

[30] Philip S Brenner and John D DeLamater. Social desirability bias in self-reports of physical activity: Is an exercise identity the culprit? *Social Indicators Research*, 117(2):489–504, 2014.

[31] Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)*, 46(3):33, 2014.

[32] Luca Canzian and Mirco Musolesi. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 1293–1304. ACM, 2015.

[33] Bokai Cao, Lei Zheng, Chenwei Zhang, Philip S Yu, Andrea Piscitello, John Zulueta, Olu Ajilore, Kelly Ryan, and Alex D Leow. Deepmood: Modeling mobile phone typing dynamics for mood detection. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 747–755. ACM, 2017.

[34] Zhen-Bo Cao, Nobuyuki Miyatake, Mitsuru Higuchi, Motohiko Miyachi, Kazuko Ishikawa-Takata, and Izumi Tabata. Predicting $\dot{v}$ o2max with an objectively measured physical activity in japanese women. *Medicine & Science in Sports & Exercise*, 42(1):179–186, 2010.

[35] James Carpenter and John Bithell. Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in medicine*, 19(9): 1141–1164, 2000.

[36] Debaditya Chakraborty and Hazem Elzarka. Advanced machine learning techniques for building performance simulation: a comparative analysis. *Journal of Building Performance Simulation*, 12(2):193–207, 2019.

[37] Yu-Feng Yvonne Chan, Pei Wang, Linda Rogers, Nicole Tignor, Micol Zweig, Steven G Hershman, Nicholas Genes, Erick R Scott, Eric Krock, Marcus Badgeley, et al. The asthma mobile health study, a large-scale clinical observational study using researchkit. *Nature biotechnology*, 35(4):354–362, 2017.

[38] Peter H Charlton, Drew A Birrenkott, Timothy Bonnici, Marco AF Pimentel, Alistair EW Johnson, Jordi Alastruey, Lionel Tarassenko, Peter J Watkinson, Richard Beale, and David A Clifton. Breathing rate estimation from the electrocardiogram and photoplethysmogram: A review. *IEEE reviews in biomedical engineering*, 11: 2–20, 2017.

[39] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):1–12, 2018.

[40] Hugh Chen, Scott M Lundberg, Gabriel Erion, Jerry H Kim, and Su-In Lee. Forecasting adverse surgical events using self-supervised transfer learning for physiological signals. *NPJ Digital Medicine*, 4(1):1–13, 2021.

[41] Richard Chen, Filip Jankovic, Nikki Marinsek, Luca Foschini, Lampros Kourtis, Alessio Signorini, Melissa Pugh, Jie Shen, Roy Yaari, Vera Maljkovic, et al. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In *KDD*, 2019.

[42] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *KDD*, 2016.

[43] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[44] Helen Cheng and Adrian Furnham. Personality, self-esteem, and demographic predictions of happiness and depression. *Personality and individual differences*, 34 (6):921–942, 2003.

[45] Joseph Y Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi. Subject-aware contrastive learning for biosignals. *arXiv preprint arXiv:2007.04871*, 2020.

[46] Charles M Ching, A Timothy Church, Marcia S Katigbak, Jose Alberto S Reyes, Junko Tanaka-Matsumi, Shino Takaoka, Hengsheng Zhang, Jiliang Shen, Rina Mazuera Arias, Brigida Carolina Rincon, et al. The manifestation of traits

in everyday behavior and affect: A five-culture study. *Journal of Research in Personality*, 48:1–16, 2014.

[47] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *EMNLP*, 2014.

[48] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, 2018.

[49] Dirk L Christensen, Daniel Faurholt-Jepsen, Michael K Boit, David L Mwaniki, Beatrice Kilonzo, Inge Tetens, Festus K Kiplamai, SC Cheruiyot, Henrik Friis, Knut Borch-Johnsen, et al. Cardiorespiratory fitness and physical activity in luo, kamba, and maasai of rural kenya. *American Journal of Human Biology*, 24(6): 723–729, 2012.

[50] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[51] Kasie D Cooper and Alex B Shafer. Validity and reliability of the polar a300's fitness test feature to predict vo2max. *International journal of exercise science*, 12(4): 393, 2019.

[52] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020.

[53] Mihaly Csikszent and Reed Larson. Validity and reliability of the experience-sampling method. In *Flow and the foundations of positive psychology*, pages 35–54. Springer, 2014.

[54] Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *AAAI*, 2018.

[55] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[56] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

[57] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94, 2019.

[58] CT Davies. Limitations to the prediction of maximum oxygen intake from cardiac frequency measurements. *Journal of Applied Physiology*, 24(5):700–706, 1968.

[59] JA Davis. Direct determination of aerobic power. *Physiological assessment of human fitness*, pages 9–17, 1995.

[60] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[61] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[62] Aiden Doherty, Dan Jackson, Nils Hammerla, Thomas Plötz, Patrick Olivier, Malcolm H Granat, Tom White, Vincent T Van Hees, Michael I Trenell, and Christoper G and others Owen. Large scale population assessment of physical activity using wrist worn accelerometers: the UK Biobank study. *PloS one*, 12(2): e0169649, 2017.

[63] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015.

[64] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[65] Allied Dunbar. National fitness survey: a report on activity patterns and fitness levels. *Sports Council and Health Education Authority: London, UK*, 1992.

[66] Micah T Eades, Athanasios Tsanas, Stephen P Juraschek, Daniel B Kramer, Ernest Gervino, and Kenneth J Mukamal. Smartphone-recorded physical activity for estimating cardiorespiratory fitness. *Scientific reports*, 11(1):1–6, 2021.

[67] Lars-Göran Ekelund, William L Haskell, Jeffrey L Johnson, Fredrick S Whaley, Michael H Criqui, David S Sheps, and Lipid Research Clinics Mortality Follow

up Study. Physical fitness as a predictor of cardiovascular mortality in asymptomatic north american men. *New England Journal of Medicine*, 319(21):1379–1384, 1988.

[68] MYRVIN H Ellestad and MKE Wan. Predictive implications of stress testing. follow-up of 2700 subjects after maximum treadmill stress testing. *Circulation*, 51 (2):363–369, 1975.

[69] Michael R Esco, Emmanuel M Mugu, Henry N Williford, Aindrea N McHugh, and Barbara E Bloomquist. Cross-validation of the polar fitness test via the polar f11 heart rate monitor in predicting vo 2 max. *Journal of Exercise Physiology Online*, 14(5), 2011.

[70] Maria Faurholt-Jepsen, Søren Brage, Lars Vedel Kessing, and Klaus Munkholm. State-related differences in heart rate variability in bipolar disorder. *Journal of psychiatric research*, 84:169–173, 2017.

[71] 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. Resting heart rate in cardiovascular disease. *Journal of the American College of Cardiology*, 50(9):823–830, 2007.

[72] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[73] Ben D Fulcher. Feature-based time-series analysis. In *Feature Engineering for Machine Learning and Data Analytics*, pages 87–116. CRC Press, 2018.

[74] Ben D Fulcher and Nick S Jones. Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):3026–3037, 2014.

[75] Héléna A Gaspar and Gerome Breen. Probabilistic ancestry maps: a method to assess and visualize population substructures in genetics. *BMC bioinformatics*, 20 (1):1–11, 2019.

[76] Katharina Geukes, Steffen Nestler, Roos Hutteman, Albrecht CP Küfner, and Mitja D Back. Trait personality and state variability: Predicting individual differences in within-and cross-context fluctuations in affect, self-evaluations, and behavior in everyday life. *Journal of Research in Personality*, 69:124–138, 2017.

[77] Marzyeh Ghassemi, Marco Pimentel, Tristan Naumann, Thomas Brennan, David Clifton, Peter Szolovits, and Mengling Feng. A multivariate timeseries modeling

approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[78] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[79] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[80] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

[81] Sharon Golub and Denise M Harrington. Premenstrual and menstrual mood changes in adolescent women. *Journal of Personality and Social Psychology*, 41(5): 961, 1981.

[82] Tomas I Gonzales, Justin Y Jeon, Timothy Lindsay, Kate Westgate, Ignacio Perez-Pozuelo, Stefanie Hollidge, Katrien Wijndaele, Kirsten Rennie, Nita Forouhi, Simon Griffin, et al. Resting heart rate as a biomarker for tracking change in cardiorespiratory fitness of uk adults: the fenland study. *medRxiv*, 2020.

[83] Tomas I Gonzales, Kate Westgate, Stefanie Hollidge, Tim Lindsay, Justin Jeon, and Soren Brage. Estimating maximal oxygen consumption from heart rate response to submaximal ramped treadmill test. *medRxiv*, 2020.

[84] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

[85] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37 (6):504–528, 2003.

[86] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 273–278. IEEE, 2013.

[87] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[88] Agnes Grünerbl, Amir Muaremi, Venet Osmani, Gernot Bahle, Stefan Oehler, Gerhard Tröster, Oscar Mayora, Christian Haring, and Paul Lukowicz. Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE Journal of Biomedical and Health Informatics*, 19(1):140–148, 2015.

[89] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[90] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[91] Haraldur T Hallgrímsson, Filip Jankovic, Tim Althoff, and Luca Foschini. Learning individualized cardiovascular responses from large-scale wearable sensors data. *Machine Learning for Healthcare at NIPS*, 2018.

[92] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. Deep, convolutional, and recurrent models for human activity recognition using wearables. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 1533–1540. AAAI Press, 2016.

[93] Mo Han, Ozan Ozdenizci, Toshiaki Koike-Akino, Ye Wang, and Deniz Erdogmus. Universal physiological representation learning with soft-disentangled rateless autoencoders. *IEEE Journal of Biomedical and Health Informatics*, 2021.

[94] Yifan Hao and Huiping Cao. A new attention mechanism to classify multivariate time series. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020.

[95] Harish Haresamudram, David V Anderson, and Thomas Plötz. On the role of features in human activity recognition. In *Proceedings of the 23rd International Symposium on Wearable Computers*, pages 78–88, 2019.

[96] Charles R. Harris, K. Jarrod Millman, St'efan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fern'andez del R'10, Mark Wiebe,

Pearu Peterson, Pierre G'erard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

[97] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[98] Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

[99] André Henriksen, Martin Haugen Mikalsen, Ashenafi Zebene Woldaregay, Miroslav Muzny, Gunnar Hartvigsen, Laila Arnesdatter Hopstock, and Sameline Grimsgaard. Using fitness trackers and smartwatches to measure physical activity in research: analysis of consumer wrist-worn wearables. *Journal of medical Internet research*, 20(3):e110, 2018.

[100] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[101] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[102] Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3):364–373, 2020.

[103] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[104] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

[105] John M Jakicic, Caitlin M Egan, Anthony N Fabricatore, Sarah A Gaussoin, Stephen P Glasser, Louise A Hesson, William C Knowler, Wei Lang, Judith G Regensteiner, Paul M Ribisl, et al. Four-year change in cardiorespiratory fitness and influence on glycemic control in adults with type 2 diabetes in a randomized trial: the look ahead trial. *Diabetes care*, 36(5):1297–1303, 2013.

[106] Natasha Jaques, Ognjen (Oggi) Rudovic, Sara Taylor, Akane Sano, and Rosalind Picard. Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation. In *Proceedings of IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, volume 66 of *Proceedings of Machine Learning Research*, pages 17–33, August 2017.

[107] Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. In *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, Texas*, 2017.

[108] Natasha Jaques, Sara Taylor, Akane Sano, Rosalind Picard, et al. 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*, pages 17–33, 2017.

[109] Shayan Jawed, Josif Grabocka, and Lars Schmidt-Thieme. Self-supervised learning for semi-supervised time series classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 499–511. Springer, 2020.

[110] Simon Jenni and Paolo Favaro. Self-supervised feature learning by learning to spot artifacts. In *CVPR*, 2018.

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

[112] Radim Jurca, Andrew S Jackson, Michael J LaMonte, James R Morrow Jr, Steven N Blair, Nicholas J Wareham, William L Haskell, Willem van Mechelen, Timothy S Church, John M Jakicic, et al. Assessing cardiorespiratory fitness without performing exercise testing. *American journal of preventive medicine*, 29(3):185–193, 2005.

[113] Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017.

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

[115] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[116] Eric S Kim, Nansook Park, Jennifer K Sun, Jacqui Smith, and Christopher Peterson. Life satisfaction and frequency of doctor visits. *Psychosomatic medicine*, 76(1):86, 2014.

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

[118] Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals. *arXiv preprint arXiv:2005.13249*, 2020.

[119] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[120] Peter F Kokkinos, Charles Faselis, Jonathan Myers, Demosthenes Panagiotakos, and Michael Doumas. Interactive effects of fitness and statin treatment on mortality risk in veterans with dyslipidaemia: a cohort study. *The Lancet*, 381 (9864):394–399, 2013.

[121] Adit Krishnan, Ashish Sharma, and Hari Sundaram. 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*, pages 297–306, 2018.

[122] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[123] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317, 2015.

[124] Peter Kuppens. Individual differences in the relationship between pleasure and arousal. *Journal of Research in Personality*, 42(4):1053–1059, 2008.

[125] Peter Kuppens, Francis Tuerlinckx, James A Russell, and Lisa Feldman Barrett. The relation between valence and arousal in subjective experience. *Psychological Bulletin*, 139(4):917, 2013.

[126] Soon Bin Kwon, Joong Woo Ahn, Seung Min Lee, Joonnyong Lee, Dongheon Lee, Jeeyoung Hong, Hee Chan Kim, and Hyung-Jin Yoon. Estimating maximal

oxygen uptake from daily activity data measured by a watch-type fitness tracker: cross-sectional study. *JMIR mHealth and uHealth*, 7(6):e13327, 2019.

[127] Susan M Labott, Timothy P Johnson, Michael Fendrich, and Norah C Feeny. Emotional risks to respondents in survey research: Some empirical evidence. *Journal of Empirical Research on Human Research Ethics*, 8(4):53–66, 2013.

[128] Timo A Lakka, Juha M Venalainen, Rainer Rauramaa, Riitta Salonen, Jaakko Tuomilehto, and Jukka T Salonen. Relation of leisure-time physical activity and cardiorespiratory fitness to the risk of acute myocardial infarction in men. *New England Journal of Medicine*, 330(22):1549–1554, 1994.

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

[130] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. A survey of mobile phone sensing. *IEEE Communications magazine*, 48(9), 2010.

[131] Nicholas D. Lane, Mashfiqui Mohammod, Mu Lin, Xiaochao Yang, Hong Lu, Shahid Ali, Afsaneh Doryab, Ethan Berke, Tanzeem Choudhury, and Andrew T. Campbell. Bewell: A smartphone application to monitor, model and promote wellbeing. In *Pervasive Computing Technologies for Healthcare*, 2011.

[132] Neal Lathia, Gillian M Sandstrom, Cecilia Mascolo, and Peter J Rentfrow. Happier people live more active lives: using smartphones to link happiness and physical activity. *PloS one*, 12(1):e0160589, 2017.

[133] Jari A Laukkanen, Sudhir Kurl, Riitta Salonen, Rainer Rauramaa, and Jukka T Salonen. The predictive value of cardiorespiratory fitness for cardiovascular events in men with various risk profiles: a prospective population-based cohort study. *European heart journal*, 25(16):1428–1437, 2004.

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

[135] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436, 2015.

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

[137] Kyunghee Lee, Hyeyon Kwon, Byungtae Lee, Guna Lee, Jae Ho Lee, Yu Rang Park, and Soo-Yong Shin. Effect of self-monitoring on long-term patient engagement with mobile health applications. *PloS one*, 13(7):e0201166, 2018.

[138] Boning Li and Akane Sano. Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(2):1–26, 2020.

[139] Da Li and Timothy Hospedales. Online meta-learning for multi-source and semi-supervised domain adaptation. In *European Conference on Computer Vision*, pages 382–403. Springer, 2020.

[140] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[141] Robert LiKamWa, Yunxin Liu, Nicholas D. Lane, and Lin Zhong. Moodscope: Building a mood sensor from smartphone usage patterns. In *MobiSys '13*. ACM, 2013.

[142] Tim Lindsay, Kate Westgate, Katrien Wijndaele, Stefanie Hollidge, Nicola Kerrison, Nita Forouhi, Simon Griffin, Nick Wareham, and Søren Brage. Descriptive epidemiology of physical activity energy expenditure in uk adults (the fenland study). *International Journal of Behavioral Nutrition and Physical Activity*, 16(1):1–13, 2019.

[143] Seppo Linnainmaa. The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors. *Master's Thesis (in Finnish), Univ. Helsinki*, pages 6–7, 1970.

[144] Terrance Liu, Paul Pu Liang, Michal Muszynski, Ryo Ishii, David Brent, Randy Auerbach, Nicholas Allen, and Louis-Philippe Morency. Multimodal privacy-preserving mood prediction from mobile data: A preliminary study. *arXiv preprint arXiv:2012.02359*, 2020.

[145] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218*, 1(2), 2020.

[146] Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A

deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, 26(6):900–908, 2020.

[147] Donald M Lloyd-Jones, Yuling Hong, Darwin Labarthe, Dariush Mozaffarian, Lawrence J Appel, Linda Van Horn, Kurt Greenlund, Stephen Daniels, Graham Nichol, Gordon F Tomaselli, et al. Defining and setting national goals for cardiovascular health promotion and disease reduction: the american heart association's strategic impact goal through 2020 and beyond. *Circulation*, 121(4): 586–613, 2010.

[148] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[149] Alicia Lozano-Diez, Ruben Zazo, Doroteo T Toledano, and Joaquin Gonzalez-Rodriguez. An analysis of the influence of deep neural network (dnn) topology in bottleneck feature based language recognition. *PloS one*, 12(8):e0182580, 2017.

[150] Naomi D Lucio, Elvia V Salazar, Ivan A Figueroa, Jose L Gamez, Ryan D Russell, and Merrill D Funk. Accuracy of fitbit charge 2 at estimating vo2max, calories, and steps on a treadmill. In *International Journal of Exercise Science: Conference Proceedings*, volume 2, page 11, 2018.

[151] John Lynch, Susan P Helmrich, Timo A Lakka, George A Kaplan, Richard D Cohen, Riitta Salonen, and Jukka T Salonen. Moderately intense physical activities and high levels of cardiorespiratory fitness reduce the risk of non-insulin-dependent diabetes mellitus in middle-aged men. *Archives of internal medicine*, 156(12):1307–1314, 1996.

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

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

[154] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

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

[156] 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. Daily activity energy expenditure and mortality among older adults. *Jama*, 296(2):171–179, 2006.

[157] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

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

[159] Caroline Marra, Jacqueline L Chen, Andrea Coravos, and Ariel D Stern. Quantifying the use of connected digital products in clinical research. *NPJ digital medicine*, 3(1):1–5, 2020.

[160] Michael V McConnell, Anna Shcherbina, Aleksandra Pavlovic, Julian R Homburger, Rachel L Goldfeder, Daryl Waggot, Mildred K Cho, Mary E Rosenberger, William L Haskell, Jonathan Myers, et al. Feasibility of obtaining measures of lifestyle from a smartphone app: the myheart counts cardiovascular health study. *JAMA cardiology*, 2(1):67–76, 2017.

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

[162] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[163] Abhinav Mehrotra, Jo Vermeulen, Veljko Pejovic, and Mirco Musolesi. Ask, but don't interrupt: the case for interruptibility-aware mobile experience sampling. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pages 723–732. ACM, 2015.

[164] Mehdi Menai, Vincent T Van Hees, Alexis Elbaz, Mika Kivimaki, Archana Singh-Manoux, and Séverine Sabia. Accelerometer assessed moderate-to-vigorous

physical activity and successful ageing: results from the whitehall ii study. *Scientific reports*, 7:45772, 2017.

[165] Gatis Mikelsons, Matthew Smith, Abhinav Mehrotra, and Mirco Musolesi. Towards deep learning models for psychological state prediction using smartphone data: Challenges and opportunities. In *In Workshop on Machine Learning for Health (ML4H) at NIPS 2017*, December 2017.

[166] David C Mohr, Mi Zhang, and Stephen M Schueller. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology*, 13:23–47, 2017.

[167] David C Mohr, Katie Shilton, and Matthew Hotopf. Digital phenotyping, behavioral sensing, or personal sensing: names and transparency in the digital age. *NPJ digital medicine*, 3(1):1–2, 2020.

[168] Hans Moravec. When will computer hardware match the human brain. *Journal of evolution and technology*, 1(1):10, 1998.

[169] Jonathan Myers, Manish Prakash, Victor Froelicher, Dat Do, Sara Partington, and J Edwin Atwood. Exercise capacity and mortality among men referred for exercise testing. *New England journal of medicine*, 346(11):793–801, 2002.

[170] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, 2016.

[171] Bjarne Martens Nes, Imre Janszky, Lars Johan Vatten, Tom Ivar Lund Nilsen, Stian Thoresen Aspenes, and Ulrik Wisløff. Estimating v˙ o2peak from a nonexercise prediction model: the hunt study, norway. *Medicine & Science in Sports & Exercise*, 43(11):2024–2030, 2011.

[172] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[173] Jianmo Ni, Larry Muhlstein, and Julian McAuley. Modeling heart rate and activity data for personalized fitness recommendation. In *WWW*, 2019.

[174] Vanessa Noonan and Elizabeth Dean. Submaximal exercise testing: clinical application and interpretation. *Physical therapy*, 80(8):782–807, 2000.

[175] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[176] All of Us Research Program Investigators. The "all of us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019.

[177] Christopher Olah. Understanding lstm networks. 2015.

[178] RS Olson, W Cava, Z Mustahsan, A Varik, and JH Moore. Data-driven advice for applying machine learning to bioinformatics problems. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 23, pages 192–203, 2018.

[179] Christina Orphanidou, Timothy Bonnici, Peter Charlton, David Clifton, David Vallance, and Lionel Tarassenko. Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring. *IEEE journal of biomedical and health informatics*, 19(3):832–838, 2014.

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

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

[182] Niclas Palmius, Althanasios Tsanas, Kate EA Saunders, Amy C Bilderbeck, John R Geddes, Guy M Goodwin, and Maarten De Vos. Detecting bipolar depression from geographic location data. *IEEE Transactions on Biomedical Engineering*, 64(8): 1761–1771, 2016.

[183] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[184] Minha Park. Error analysis and stochastic modeling of mems based inertial sensors for land vehicle navigation applications. 2004.

[185] Stefanie Passler, Julian Bohrer, Lukas Blöchinger, and Veit Senner. Validity of wrist-worn activity trackers for estimating vo2max and energy expenditure. *International journal of environmental research and public health*, 16(17):3037, 2019.

[186] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[187] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2 (11):559–572, 1901.

[188] MD Peggy Bui. Using ai to help find answers to common skin conditions, May 2021. URL https://blog.google/technology/health/derm-assist-preview-io-2021.

[189] Marco V Perez, Kenneth W Mahaffey, Haley Hedlin, John S Rumsfeld, Ariadna Garcia, Todd Ferris, Vidhya Balasubramanian, Andrea M Russo, Amol Rajmane, Lauren Cheung, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. *New England Journal of Medicine*, 381(20):1909–1917, 2019.

[190] Ignacio Perez-Pozuelo, Dimitris Spathis, Emma AD Clifton, and Cecilia Mascolo. Wearables, smartphones, and artificial intelligence for digital phenotyping and health. In *Digital Health*, pages 33–54. Elsevier.

[191] Ignacio Perez-Pozuelo, Dimitris Spathis, Jordan Gifford-Moore, Jessica Morley, and Josh Cowls. Digital phenotyping and sensitive health data: Implications for data governance. *Journal of the American Medical Informatics Association*, 2021.

[192] Lerrel Pinto and Abhinav Gupta. Learning to push by grasping: Using multiple tasks for effective learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 2161–2168. IEEE, 2017.

[193] Guy Plasqui and Klaas R Westerterp. Accelerometry and heart rate as a measure of physical fitness: proof of concept. *Medicine & Science in Sports & Exercise*, 37 (5):872–876, 2005.

[194] Jennifer M Radin, Nathan E Wineinger, Eric J Topol, and Steven R Steinhubl. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the usa: a population-based study. *The Lancet Digital Health*, 2(2):e85–e93, 2020.

[195] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):157, 2018.

[196] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6): 96–108, 2017.

[197] Aadia I Rana, Jacob J van den Berg, Eric Lamy, and Curt G Beckwith. Using a mobile health intervention to support hiv treatment adherence and retention among patients at risk for disengaging with care. *AIDS patient care and STDs*, 30 (4):178–184, 2016.

[198] Ruth Reader. 2 stanford experts say ai won't transform healthcare until the 2030s, Apr 2021. URL https://www.fastcompany.com/90630654/stanford-ai-experts-healthcare.

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

[200] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[201] Robert Ross, Steven N Blair, Ross Arena, Timothy S Church, Jean-Pierre Després, Barry A Franklin, William L Haskell, Leonard A Kaminsky, Benjamin D Levine, Carl J Lavie, et al. Importance of assessing cardiorespiratory fitness in clinical practice: a case for fitness as a clinical vital sign: a scientific statement from the american heart association. *Circulation*, 134(24):e653–e699, 2016.

[202] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[203] Yulia Rubanova, Ricky TQ Chen, and David Duvenaud. Latent odes for irregularly-sampled time series. *arXiv preprint arXiv:1907.03907*, 2019.

[204] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[205] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[206] James A. Russell, Anna Weiss, and Gerald A. Mendelsohn. Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, 1989.

[207] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. Multi-task self-supervised learning for human activity detection. *IMWUT*, 3(2):61, 2019.

[208] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[209] James F Sallis and Brian E Saelens. Assessment of physical activity by self-report: status, limitations, and future directions. *Research quarterly for exercise and sport*, 71(sup2):1–14, 2000.

[210] Benjamin Sanchez-Lengeling, Jennifer N Wei, Brian K Lee, Richard C Gerkin, Alán Aspuru-Guzik, and Alexander B Wiltschko. Machine learning for scent: Learning generalizable perceptual representations of small molecules. *arXiv preprint arXiv:1910.10685*, 2019.

[211] Akane Sano. *Measuring college students' sleep, stress, mental health and wellbeing with wearable sensors and mobile phones*. PhD thesis, Massachusetts Institute of Technology, 2016.

[212] Akane Sano, Z Yu Amy, Andrew W McHill, Andrew JK Phillips, Sara Taylor, Natasha Jaques, Elizabeth B Klerman, and Rosalind W Picard. Prediction of happy-sad mood from daily behaviors and previous sleep history. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 6796–6799. IEEE, 2015.

[213] Pritam Sarkar and Ali Etemad. Self-supervised learning for ecg-based emotion recognition. *arXiv preprint arXiv:1910.07497*, 2019.

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

[215] D Schmid and MF Leitzmann. Cardiorespiratory fitness as predictor of cancer mortality: a systematic review and meta-analysis. *Annals of oncology*, 26(2): 272–278, 2015.

[216] Dale A Schoeller. Limitations in the assessment of dietary energy intake by self-report. *Metabolism*, 44:18–22, 1995.

[217] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[218] Charlotte Schubert, Gareth Archer, Jo M Zelis, Sarah Nordmeyer, Kilian Runte, Anja Hennemuth, Felix Berger, Volkmar Falk, Pim AL Tonino, Rod Hose, et al. Wearable devices can predict the outcome of standardized 6-minute walk tests in heart disease. *NPJ digital medicine*, 3(1):1–9, 2020.

[219] Felipe B Schuch, Davy Vancampfort, Xuemei Sui, Simon Rosenbaum, Joseph Firth, Justin Richards, Philip B Ward, and Brendon Stubbs. Are lower levels of cardiorespiratory fitness associated with incident depression? a systematic review of prospective cohort studies. *Preventive Medicine*, 93:159–165, 2016.

[220] Patrick Schwab and Walter Karlen. Phonemd: Learning to diagnose parkinson's disease from smartphone data. In *AAAI*, 2019.

[221] Katrina J Serrano, Kisha I Coa, Mandi Yu, Dana L Wolff-Hughes, and Audie A Atienza. Characterizing user engagement with health app data: a data mining approach. *Translational behavioral medicine*, 7(2):277–285, 2017.

[222] Sandra Servia-Rodríguez, Kiran K Rachuri, Cecilia Mascolo, Peter J Rentfrow, Neal Lathia, and Gillian M Sandstrom. Mobile sensing at the service of mental well-being: a large-scale longitudinal study. In *Proceedings of the 26th International Conference on World Wide Web*, pages 103–112. International World Wide Web Conferences Steering Committee, 2017.

[223] Rutvik V Shah, Gillian Grennan, Mariam Zafar-Khan, Fahad Alim, Sujit Dey, Dhakshin Ramanathan, and Jyoti Mishra. Personalized machine learning of depressed mood using wearables. *Translational Psychiatry*, 11(1):1–18, 2021.

[224] Anna Shcherbina, C Mikael Mattsson, Daryl Waggott, Heidi Salisbury, Jeffrey W Christle, Trevor Hastie, Matthew T Wheeler, and Euan A Ashley. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *Journal of personalized medicine*, 7(2):3, 2017.

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

[226] Dimitris Spathis, Sandra Servia-Rodriguez, Katayoun Farrahi, Cecilia Mascolo, and Jason Rentfrow. 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*, pages 272–281, 2019.

[227] Dimitris Spathis, Sandra Servia-Rodriguez, Katayoun Farrahi, Cecilia Mascolo, and Jason Rentfrow. 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*, pages 2886–2894, 2019.

[228] Dimitris Spathis, Ignacio Perez-Pozuelo, Soren Brage, Nicholas J Wareham, and Cecilia Mascolo. Learning generalizable physiological representations from large-scale wearable data. *arXiv preprint arXiv:2011.04601*, 2020.

[229] Dimitris Spathis, Ignacio Perez-Pozuelo, Soren Brage, Nicholas J. Wareham, and Cecilia Mascolo. Self-supervised transfer learning of physiological representations from free-living wearable data. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, page 69–78, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383592. doi: 10.1145/3450439.3451863. URL https://doi.org/10.1145/3450439.3451863.

[230] Dimitris Spathis, Ignacio Perez-Pozuelo, Laia Marques-Fernandez, and Cecilia Mascolo. Breaking away from labels: the promise of self-supervised machine learning in intelligent health. *Patterns (in press)*, 2022.

[231] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[232] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.

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

[234] Yoshihiko Suhara, Yinzhan Xu, and Alex'Sandy' Pentland. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *Proceedings of the 26th International Conference on World Wide Web*,

pages 715–724. International World Wide Web Conferences Steering Committee, 2017.

[235] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[236] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[237] David P Swain, Clinton A Brawner, American College of Sports Medicine, et al. *ACSM's resource manual for guidelines for exercise testing and prescription*. Wolters Kluwer Health/Lippincott Williams & Wilkins, 2014.

[238] Ann M Swank, John Horton, Jerome L Fleg, Gregg C Fonarow, Steven Keteyian, Lee Goldberg, Gene Wolfel, Eileen M Handberg, Dan Bensimhon, Marie-Christine Illiou, et al. Modest increase in peak vo2 is related to better clinical outcomes in chronic heart failure patients: results from heart failure and a controlled trial to investigate outcomes of exercise training. *Circulation: Heart Failure*, 5(5):579–585, 2012.

[239] Setareh Rahimi Taghanaki and Ali Etemad. Self-supervised wearable-based activity recognition by learning to forecast motion. *arXiv preprint arXiv:2010.13713*, 2020.

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

[241] Matthew Tancik, Pratul P Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *arXiv preprint arXiv:2006.10739*, 2020.

[242] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, and Cecilia Mascolo. Exploring contrastive learning in human activity recognition for healthcare. *arXiv preprint arXiv:2011.11542*, 2020.

[243] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, Soren Brage, Nick Wareham, and Cecilia Mascolo. Selfhar: Improving human activity recognition through self-training with unlabeled data. *arXiv preprint arXiv:2102.06073*, 2021.

[244] Sara Ann Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing*, 2017.

[245] Nabyl Tejani, Timothy R Dresselhaus, and Matthew B Weinger. Development of a hand-held computer platform for real-time behavioral assessment of physicians and nurses. *Journal of biomedical informatics*, 43(1):75–80, 2010.

[246] Anja Thieme, Danielle Belgrave, and Gavin Doherty. Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(5):1–53, 2020.

[247] Robert L Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[248] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision, 2021.

[249] Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. *arXiv preprint arXiv:2106.00750*, 2021.

[250] Catherine Tong, Shyam A Tailor, and Nicholas D Lane. Are accelerometers for activity recognition a dead-end? In *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications*, pages 39–44, 2020.

[251] Gautam K Vallabha, James L McClelland, Ferran Pons, Janet F Werker, and Shigeaki Amano. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, 104(33):13273–13278, 2007.

[252] Niels Van Berkel, Denzil Ferreira, and Vassilis Kostakos. The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR)*, 50(6):93, 2018.

[253] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR abs/1609.03499*, 2016.

[254] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[255] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017.

[256] Ruut Veenhoven. Healthy happiness: Effects of happiness on physical health and the consequences for preventive health care. *Journal of happiness studies*, 9(3): 449–469, 2008.

[257] Arun Venkatraman, Martial Hebert, and J Andrew Bagnell. Improving multi-step prediction of learned time series models. In *AAAI*, pages 3024–3030, 2015.

[258] Alex Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3):328–339, 1989.

[259] Anran Wang, Jiwen Lu, Jianfei Cai, Tat-Jen Cham, and Gang Wang. Large-margin multi-modal deep learning for rgb-d object recognition. *IEEE Transactions on Multimedia*, 17(11):1887–1898, 2015.

[260] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Wenjun Zeng, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. *arXiv preprint arXiv:2103.03097*, 2021.

[261] Luyu Wang, Pauline Luc, Adria Recasens, Jean-Baptiste Alayrac, and Aaron van den Oord. Multimodal self-supervised learning of general audio representations. *arXiv preprint arXiv:2104.12807*, 2021.

[262] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 3–14. ACM, 2014.

[263] Zhiwei Wang, Tyler Derr, Dawei Yin, and Jiliang Tang. Understanding and predicting weight loss with mobile social networking data. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1269–1278, 2017.

[264] Nicholas J Wareham, Susie J Hennings, Andrew M Prentice, and Nicholas E Day. Feasibility of heart-rate monitoring to estimate total level and pattern of energy expenditure in a population-based epidemiological study: the ely young cohort feasibility study 1994–5. *British Journal of Nutrition*, 78(6):889–900, 1997.

[265] David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

[266] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 1(10):e2, 2016.

[267] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018.

[268] Iris MR Weller, Scott G Thomas, Norm Gledhill, Don Paterson, and Art Quinney. A study to validate the modified canadian aerobic fitness test. *Canadian Journal of Applied Physiology*, 20(2):211–221, 1995.

[269] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.

[270] Tom White, Kate Westgate, Nicholas J Wareham, and Soren Brage. Estimation of physical activity energy expenditure during free-living from wrist accelerometry in uk adults. *PLoS One*, 11(12):e0167472, 2016.

[271] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Taylan Cemgil, et al. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*, 2021.

[272] Christopher S Wood, Michael R Thomas, Jobie Budd, Tivani P Mashamba-Thompson, Kobus Herbst, Deenan Pillay, Rosanna W Peeling, Anne M Johnson, Rachel A McKendry, and Molly M Stevens. Taking connected mobile-health diagnostics of infectious diseases to the field. *Nature*, 566(7745):467–474, 2019.

[273] World Health Organization. *The world health report 2002: reducing risks, promoting healthy life*. World Health Organization, 2002.

[274] Xian Wu, Chao Huang, Pablo Roblesgranda, and Nitesh Chawla. Representation learning on variable length and incomplete wearable-sensory time series. *arXiv preprint arXiv:2002.03595*, 2020.

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

[276] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020.

[277] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th International Conference on World Wide Web*, pages 351–360. International World Wide Web Conferences Steering Committee, 2017.

[278] Hugo Yèche, Gideon Dresdner, Francesco Locatello, Matthias Hüser, and Gunnar Rätsch. Neighborhood contrastive learning applied to online patient monitoring. *arXiv preprint arXiv:2106.05142*, 2021.

[279] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[280] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[281] XianXing Zhang, Yitong Zhou, Yiming Ma, Bee-Chung Chen, Liang Zhang, and Deepak Agarwal. Glmix: Generalized linear mixed models for large-scale response prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 363–372. ACM, 2016.

[282] Xiao Zhang, Wenzhong Li, Xu Chen, and Sanglu Lu. Moodexplorer: Towards compound emotion detection via smartphone sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):176, 2018.

Appendix A

Extra information

A.1 Hyperparameters

Here we list the hyperparameters used to fine-tune the models proposed throughout the thesis.

A.1.1 Chapter 3 model

For the MLP network, we ended up with a 3-dense layer architecture with 100-50-100 units, after a grid search of {1-3, with 3 options} layer depth and unit dimensionality of {25-100, with 4 options}. There was an initial heuristic search in order to narrow down these ranges. The dropout rate was found through a search of {0.25-0.75, with 3 options}. For the rest of the baselines, the sklearn hyperparameters were used.

A.1.2 Chapter 4 model

For the encoder-decoder LSTM, we endep up with an architecture of 80-80 units, after a grid search of unit dimensionality {20-100, 5 options}. There was an initial heuristic search in order to narrow down these ranges. The recurrent dropout rate was found through a search of {0.25-0.75, with 3 options}. For the rest of the baselines, the sklearn hyperparameters were used.

A.1.3 Chapter 5 model

For the multimodal CNN, we ended up with an architecture of two 128d CNNs followed by two 128d Bi-GRUs and a pooling layer. The other two branches of the network used two 128d Dense layers followed by a dropout layer. The last Dense layer concatenates the outputs [256, 128, 128] of all branches and predicts a single value. The

unit dimensions were found through a grid search of units {64-256, 3 options}, and dropout through a search of {0.3-0.66, 3 options}. The XGboost hyperparameters were found through 5-fold cross validation using a learning rate of {0.05-0.30, 6 options}.

For the transfer learning task, a Logistic Regression classifier with a balanced class weight was trained with a 5-fold cross validation and the best model was applied to the test set. The best hyperparameters were found through a random search of 20 iterations picking from a uniform distribution with 0 lower bound and the distribution range of 10, for the inverse of regularization strength parameter (”C”), and the norm of the penalty between {L2, L1}.

A.1.4 Chapter 6 model

For the main neural network, we ended up with an architecture of two 128d Dense layers, whose dimensions were found after grid search of {64-256, 3 options}, dropout through a search of {0.3-0.6, 3 options}, layer depth of {1-3, with 3 options}, and the activation is ’elu’ which was found through a search of {’elu’, ’relu’}. For the network in Task 2, the binary task has one 128d Dense layer followed by a 0.5 Dropout found after searching the same parameters.

A.2 Feature list

Below is the full feature list for the models used in Chapter 3. We note that this feature set was extracted for each of the sensor modalities. For a detailed overview of the naming conventions we refer the reader to the documentation of the tsfresh library.

```
[abs_energy',
 absolute_sum_of_changes',
 agg_autocorrelation__f_agg_"mean"',
 agg_autocorrelation__f_agg_"median"',
 agg_autocorrelation__f_agg_"var"',
 agg_linear_trend__f_agg_"max"__chunk_len_10__attr_"intercept"',
 agg_linear_trend__f_agg_"max"__chunk_len_10__attr_"rvalue"',
 agg_linear_trend__f_agg_"max"__chunk_len_10__attr_"slope"',
 agg_linear_trend__f_agg_"max"__chunk_len_10__attr_"stderr"',
 agg_linear_trend__f_agg_"max"__chunk_len_5__attr_"intercept"',
 agg_linear_trend__f_agg_"max"__chunk_len_5__attr_"rvalue"',
 agg_linear_trend__f_agg_"max"__chunk_len_5__attr_"slope"',
 agg_linear_trend__f_agg_"max"__chunk_len_5__attr_"stderr"',
 agg_linear_trend__f_agg_"mean"__chunk_len_10__attr_"intercept"',
 agg_linear_trend__f_agg_"mean"__chunk_len_10__attr_"rvalue"',
 agg_linear_trend__f_agg_"mean"__chunk_len_10__attr_"slope"',
 agg_linear_trend__f_agg_"mean"__chunk_len_10__attr_"stderr"',
```

```
agg_linear_trend__f_agg_"mean"__chunk_len_5__attr_"intercept"’,
agg_linear_trend__f_agg_"mean"__chunk_len_5__attr_"rvalue"’,
agg_linear_trend__f_agg_"mean"__chunk_len_5__attr_"slope"’,
agg_linear_trend__f_agg_"mean"__chunk_len_5__attr_"stderr"’,
agg_linear_trend__f_agg_"min"__chunk_len_10__attr_"intercept"’,
agg_linear_trend__f_agg_"min"__chunk_len_10__attr_"rvalue"’,
agg_linear_trend__f_agg_"min"__chunk_len_10__attr_"slope"’,
agg_linear_trend__f_agg_"min"__chunk_len_10__attr_"stderr"’,
agg_linear_trend__f_agg_"min"__chunk_len_5__attr_"intercept"’,
agg_linear_trend__f_agg_"min"__chunk_len_5__attr_"rvalue"’,
agg_linear_trend__f_agg_"min"__chunk_len_5__attr_"slope"’,
agg_linear_trend__f_agg_"min"__chunk_len_5__attr_"stderr"’,
agg_linear_trend__f_agg_"var"__chunk_len_10__attr_"intercept"’,
agg_linear_trend__f_agg_"var"__chunk_len_10__attr_"rvalue"’,
agg_linear_trend__f_agg_"var"__chunk_len_10__attr_"slope"’,
agg_linear_trend__f_agg_"var"__chunk_len_10__attr_"stderr"’,
agg_linear_trend__f_agg_"var"__chunk_len_5__attr_"intercept"’,
agg_linear_trend__f_agg_"var"__chunk_len_5__attr_"rvalue"’,
agg_linear_trend__f_agg_"var"__chunk_len_5__attr_"slope"’,
agg_linear_trend__f_agg_"var"__chunk_len_5__attr_"stderr"’,
approximate_entropy__m_2__r_0.1’,
approximate_entropy__m_2__r_0.3’,
approximate_entropy__m_2__r_0.5’,
approximate_entropy__m_2__r_0.7’,
approximate_entropy__m_2__r_0.9’,
ar_coefficient__k_10__coeff_0’,
ar_coefficient__k_10__coeff_1’,
ar_coefficient__k_10__coeff_2’,
ar_coefficient__k_10__coeff_3’,
ar_coefficient__k_10__coeff_4’,
augmented_dickey_fuller__attr_"pvalue"’,
augmented_dickey_fuller__attr_"teststat"’,
augmented_dickey_fuller__attr_"usedlag"’,
autocorrelation__lag_0’,
autocorrelation__lag_1’,
autocorrelation__lag_2’,
autocorrelation__lag_3’,
autocorrelation__lag_4’,
autocorrelation__lag_5’,
autocorrelation__lag_6’,
autocorrelation__lag_7’,
autocorrelation__lag_8’,
autocorrelation__lag_9’,
binned_entropy__max_bins_10’,
c3__lag_1’,
c3__lag_2’,
```

```
c3__lag_3',
change_quantiles__f_agg_"mean"__isabs_False__qh_0.2__ql_0.0',
change_quantiles__f_agg_"mean"__isabs_False__qh_0.4__ql_0.0',
change_quantiles__f_agg_"mean"__isabs_False__qh_0.4__ql_0.2',
change_quantiles__f_agg_"mean"__isabs_False__qh_0.6__ql_0.0',
change_quantiles__f_agg_"mean"__isabs_False__qh_0.6__ql_0.2',
change_quantiles__f_agg_"mean"__isabs_False__qh_0.6__ql_0.4',
change_quantiles__f_agg_"mean"__isabs_False__qh_0.8__ql_0.0',
change_quantiles__f_agg_"mean"__isabs_False__qh_0.8__ql_0.2',
change_quantiles__f_agg_"mean"__isabs_False__qh_0.8__ql_0.4',
change_quantiles__f_agg_"mean"__isabs_False__qh_0.8__ql_0.6',
change_quantiles__f_agg_"mean"__isabs_False__qh_1.0__ql_0.0',
change_quantiles__f_agg_"mean"__isabs_False__qh_1.0__ql_0.2',
change_quantiles__f_agg_"mean"__isabs_False__qh_1.0__ql_0.4',
change_quantiles__f_agg_"mean"__isabs_False__qh_1.0__ql_0.6',
change_quantiles__f_agg_"mean"__isabs_False__qh_1.0__ql_0.8',
change_quantiles__f_agg_"mean"__isabs_True__qh_0.2__ql_0.0',
change_quantiles__f_agg_"mean"__isabs_True__qh_0.4__ql_0.0',
change_quantiles__f_agg_"mean"__isabs_True__qh_0.4__ql_0.2',
change_quantiles__f_agg_"mean"__isabs_True__qh_0.6__ql_0.0',
change_quantiles__f_agg_"mean"__isabs_True__qh_0.6__ql_0.2',
change_quantiles__f_agg_"mean"__isabs_True__qh_0.6__ql_0.4',
change_quantiles__f_agg_"mean"__isabs_True__qh_0.8__ql_0.0',
change_quantiles__f_agg_"mean"__isabs_True__qh_0.8__ql_0.2',
change_quantiles__f_agg_"mean"__isabs_True__qh_0.8__ql_0.4',
change_quantiles__f_agg_"mean"__isabs_True__qh_0.8__ql_0.6',
change_quantiles__f_agg_"mean"__isabs_True__qh_1.0__ql_0.0',
change_quantiles__f_agg_"mean"__isabs_True__qh_1.0__ql_0.2',
change_quantiles__f_agg_"mean"__isabs_True__qh_1.0__ql_0.4',
change_quantiles__f_agg_"mean"__isabs_True__qh_1.0__ql_0.6',
change_quantiles__f_agg_"mean"__isabs_True__qh_1.0__ql_0.8',
change_quantiles__f_agg_"var"__isabs_False__qh_0.2__ql_0.0',
change_quantiles__f_agg_"var"__isabs_False__qh_0.4__ql_0.0',
change_quantiles__f_agg_"var"__isabs_False__qh_0.4__ql_0.2',
change_quantiles__f_agg_"var"__isabs_False__qh_0.6__ql_0.0',
change_quantiles__f_agg_"var"__isabs_False__qh_0.6__ql_0.2',
change_quantiles__f_agg_"var"__isabs_False__qh_0.6__ql_0.4',
change_quantiles__f_agg_"var"__isabs_False__qh_0.8__ql_0.0',
change_quantiles__f_agg_"var"__isabs_False__qh_0.8__ql_0.2',
change_quantiles__f_agg_"var"__isabs_False__qh_0.8__ql_0.4',
change_quantiles__f_agg_"var"__isabs_False__qh_0.8__ql_0.6',
change_quantiles__f_agg_"var"__isabs_False__qh_1.0__ql_0.0',
change_quantiles__f_agg_"var"__isabs_False__qh_1.0__ql_0.2',
change_quantiles__f_agg_"var"__isabs_False__qh_1.0__ql_0.4',
change_quantiles__f_agg_"var"__isabs_False__qh_1.0__ql_0.6',
change_quantiles__f_agg_"var"__isabs_False__qh_1.0__ql_0.8',
```

```
change_quantiles__f_agg_"var"__isabs_True__qh_0.2__ql_0.0',
change_quantiles__f_agg_"var"__isabs_True__qh_0.4__ql_0.0',
change_quantiles__f_agg_"var"__isabs_True__qh_0.4__ql_0.2',
change_quantiles__f_agg_"var"__isabs_True__qh_0.6__ql_0.0',
change_quantiles__f_agg_"var"__isabs_True__qh_0.6__ql_0.2',
change_quantiles__f_agg_"var"__isabs_True__qh_0.6__ql_0.4',
change_quantiles__f_agg_"var"__isabs_True__qh_0.8__ql_0.0',
change_quantiles__f_agg_"var"__isabs_True__qh_0.8__ql_0.2',
change_quantiles__f_agg_"var"__isabs_True__qh_0.8__ql_0.4',
change_quantiles__f_agg_"var"__isabs_True__qh_0.8__ql_0.6',
change_quantiles__f_agg_"var"__isabs_True__qh_1.0__ql_0.0',
change_quantiles__f_agg_"var"__isabs_True__qh_1.0__ql_0.2',
change_quantiles__f_agg_"var"__isabs_True__qh_1.0__ql_0.4',
change_quantiles__f_agg_"var"__isabs_True__qh_1.0__ql_0.6',
change_quantiles__f_agg_"var"__isabs_True__qh_1.0__ql_0.8',
cid_ce__normalize_False',
cid_ce__normalize_True',
count_above_mean',
count_below_mean',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_0__w_10',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_0__w_2',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_0__w_20',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_0__w_5',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_10__w_10',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_10__w_2',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_10__w_20',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_10__w_5',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_11__w_10',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_11__w_2',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_11__w_20',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_11__w_5',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_12__w_10',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_12__w_2',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_12__w_20',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_12__w_5',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_13__w_10',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_13__w_2',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_13__w_20',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_13__w_5',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_14__w_10',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_14__w_2',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_14__w_20',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_14__w_5',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_1__w_10',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_1__w_2',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_1__w_20',
```

```
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_1__w_5',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_2__w_10',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_2__w_2',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_2__w_20',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_2__w_5',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_3__w_10',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_3__w_2',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_3__w_20',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_3__w_5',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_4__w_10',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_4__w_2',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_4__w_20',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_4__w_5',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_5__w_10',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_5__w_2',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_5__w_20',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_5__w_5',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_6__w_10',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_6__w_2',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_6__w_20',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_6__w_5',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_7__w_10',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_7__w_2',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_7__w_20',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_7__w_5',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_8__w_10',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_8__w_2',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_8__w_20',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_8__w_5',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_9__w_10',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_9__w_2',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_9__w_20',
cwt_coefficients__widths_(2, 5, 10, 20)__coeff_9__w_5',
energy_ratio_by_chunks__num_segments_10__segment_focus_0',
energy_ratio_by_chunks__num_segments_10__segment_focus_1',
energy_ratio_by_chunks__num_segments_10__segment_focus_2',
energy_ratio_by_chunks__num_segments_10__segment_focus_3',
energy_ratio_by_chunks__num_segments_10__segment_focus_4',
energy_ratio_by_chunks__num_segments_10__segment_focus_5',
energy_ratio_by_chunks__num_segments_10__segment_focus_6',
energy_ratio_by_chunks__num_segments_10__segment_focus_7',
energy_ratio_by_chunks__num_segments_10__segment_focus_8',
energy_ratio_by_chunks__num_segments_10__segment_focus_9',
fft_aggregated__aggtype_"centroid"',
fft_aggregated__aggtype_"kurtosis"',
fft_aggregated__aggtype_"skew"',
```

```
fft_aggregated__aggtype_"variance"',
fft_coefficient__coeff_0__attr_"abs"',
fft_coefficient__coeff_0__attr_"real"',
fft_coefficient__coeff_10__attr_"abs"',
fft_coefficient__coeff_10__attr_"angle"',
fft_coefficient__coeff_10__attr_"imag"',
fft_coefficient__coeff_10__attr_"real"',
fft_coefficient__coeff_11__attr_"abs"',
fft_coefficient__coeff_11__attr_"angle"',
fft_coefficient__coeff_11__attr_"imag"',
fft_coefficient__coeff_11__attr_"real"',
fft_coefficient__coeff_12__attr_"abs"',
fft_coefficient__coeff_12__attr_"angle"',
fft_coefficient__coeff_12__attr_"imag"',
fft_coefficient__coeff_12__attr_"real"',
fft_coefficient__coeff_13__attr_"abs"',
fft_coefficient__coeff_13__attr_"angle"',
fft_coefficient__coeff_13__attr_"imag"',
fft_coefficient__coeff_13__attr_"real"',
fft_coefficient__coeff_14__attr_"abs"',
fft_coefficient__coeff_14__attr_"angle"',
fft_coefficient__coeff_14__attr_"imag"',
fft_coefficient__coeff_14__attr_"real"',
fft_coefficient__coeff_15__attr_"abs"',
fft_coefficient__coeff_15__attr_"angle"',
fft_coefficient__coeff_15__attr_"imag"',
fft_coefficient__coeff_15__attr_"real"',
fft_coefficient__coeff_16__attr_"abs"',
fft_coefficient__coeff_16__attr_"angle"',
fft_coefficient__coeff_16__attr_"imag"',
fft_coefficient__coeff_16__attr_"real"',
fft_coefficient__coeff_17__attr_"abs"',
fft_coefficient__coeff_17__attr_"angle"',
fft_coefficient__coeff_17__attr_"imag"',
fft_coefficient__coeff_17__attr_"real"',
fft_coefficient__coeff_18__attr_"abs"',
fft_coefficient__coeff_18__attr_"angle"',
fft_coefficient__coeff_18__attr_"imag"',
fft_coefficient__coeff_18__attr_"real"',
fft_coefficient__coeff_19__attr_"abs"',
fft_coefficient__coeff_19__attr_"angle"',
fft_coefficient__coeff_19__attr_"imag"',
fft_coefficient__coeff_19__attr_"real"',
fft_coefficient__coeff_1__attr_"abs"',
fft_coefficient__coeff_1__attr_"angle"',
fft_coefficient__coeff_1__attr_"imag"',
```

```
fft_coefficient__coeff_1__attr_"real"',
fft_coefficient__coeff_20__attr_"abs"',
fft_coefficient__coeff_20__attr_"angle"',
fft_coefficient__coeff_20__attr_"imag"',
fft_coefficient__coeff_20__attr_"real"',
fft_coefficient__coeff_21__attr_"abs"',
fft_coefficient__coeff_21__attr_"angle"',
fft_coefficient__coeff_21__attr_"imag"',
fft_coefficient__coeff_21__attr_"real"',
fft_coefficient__coeff_22__attr_"abs"',
fft_coefficient__coeff_22__attr_"angle"',
fft_coefficient__coeff_22__attr_"imag"',
fft_coefficient__coeff_22__attr_"real"',
fft_coefficient__coeff_23__attr_"abs"',
fft_coefficient__coeff_23__attr_"angle"',
fft_coefficient__coeff_23__attr_"imag"',
fft_coefficient__coeff_23__attr_"real"',
fft_coefficient__coeff_24__attr_"abs"',
fft_coefficient__coeff_24__attr_"angle"',
fft_coefficient__coeff_24__attr_"real"',
fft_coefficient__coeff_2__attr_"abs"',
fft_coefficient__coeff_2__attr_"angle"',
fft_coefficient__coeff_2__attr_"imag"',
fft_coefficient__coeff_2__attr_"real"',
fft_coefficient__coeff_3__attr_"abs"',
fft_coefficient__coeff_3__attr_"angle"',
fft_coefficient__coeff_3__attr_"imag"',
fft_coefficient__coeff_3__attr_"real"',
fft_coefficient__coeff_4__attr_"abs"',
fft_coefficient__coeff_4__attr_"angle"',
fft_coefficient__coeff_4__attr_"imag"',
fft_coefficient__coeff_4__attr_"real"',
fft_coefficient__coeff_5__attr_"abs"',
fft_coefficient__coeff_5__attr_"angle"',
fft_coefficient__coeff_5__attr_"imag"',
fft_coefficient__coeff_5__attr_"real"',
fft_coefficient__coeff_6__attr_"abs"',
fft_coefficient__coeff_6__attr_"angle"',
fft_coefficient__coeff_6__attr_"imag"',
fft_coefficient__coeff_6__attr_"real"',
fft_coefficient__coeff_7__attr_"abs"',
fft_coefficient__coeff_7__attr_"angle"',
fft_coefficient__coeff_7__attr_"imag"',
fft_coefficient__coeff_7__attr_"real"',
fft_coefficient__coeff_8__attr_"abs"',
fft_coefficient__coeff_8__attr_"angle"',
```

```
fft_coefficient__coeff_8__attr_"imag"',
fft_coefficient__coeff_8__attr_"real"',
fft_coefficient__coeff_9__attr_"abs"',
fft_coefficient__coeff_9__attr_"angle"',
fft_coefficient__coeff_9__attr_"imag"',
fft_coefficient__coeff_9__attr_"real"',
first_location_of_maximum',
first_location_of_minimum',
friedrich_coefficients__m_3__r_30__coeff_0',
friedrich_coefficients__m_3__r_30__coeff_1',
friedrich_coefficients__m_3__r_30__coeff_2',
friedrich_coefficients__m_3__r_30__coeff_3',
has_duplicate',
has_duplicate_max',
has_duplicate_min',
index_mass_quantile__q_0.1',
index_mass_quantile__q_0.2',
index_mass_quantile__q_0.3',
index_mass_quantile__q_0.4',
index_mass_quantile__q_0.6',
index_mass_quantile__q_0.7',
index_mass_quantile__q_0.8',
index_mass_quantile__q_0.9',
kurtosis',
large_standard_deviation__r_0.15000000000000002',
large_standard_deviation__r_0.2',
large_standard_deviation__r_0.25',
large_standard_deviation__r_0.30000000000000004',
large_standard_deviation__r_0.35000000000000003',
large_standard_deviation__r_0.4',
last_location_of_maximum',
last_location_of_minimum',
linear_trend__attr_"intercept"',
linear_trend__attr_"pvalue"',
linear_trend__attr_"rvalue"',
linear_trend__attr_"slope"',
linear_trend__attr_"stderr"',
longest_strike_above_mean',
longest_strike_below_mean',
max_langevin_fixed_point__m_3__r_30',
mean',
mean_abs_change',
mean_change',
mean_second_derivative_central',
median',
minimum',
```

```
number_cwt_peaks__n_1',
number_cwt_peaks__n_5',
number_peaks__n_1',
number_peaks__n_10',
number_peaks__n_3',
number_peaks__n_5',
partial_autocorrelation__lag_1',
partial_autocorrelation__lag_2',
partial_autocorrelation__lag_3',
partial_autocorrelation__lag_4',
partial_autocorrelation__lag_5',
partial_autocorrelation__lag_6',
partial_autocorrelation__lag_7',
partial_autocorrelation__lag_8',
partial_autocorrelation__lag_9',
percentage_of_reoccurring_datapoints_to_all_datapoints',
percentage_of_reoccurring_values_to_all_values',
quantile__q_0.1',
quantile__q_0.2',
quantile__q_0.3',
quantile__q_0.4',
quantile__q_0.6',
quantile__q_0.7',
quantile__q_0.8',
quantile__q_0.9',
range_count__max_1__min_-1',
ratio_beyond_r_sigma__r_0.5',
ratio_beyond_r_sigma__r_1',
ratio_beyond_r_sigma__r_1.5',
ratio_beyond_r_sigma__r_2',
ratio_beyond_r_sigma__r_2.5',
ratio_beyond_r_sigma__r_3',
ratio_beyond_r_sigma__r_5',
ratio_beyond_r_sigma__r_6',
ratio_value_number_to_time_series_length',
sample_entropy',
skewness',
spkt_welch_density__coeff_2',
spkt_welch_density__coeff_5',
spkt_welch_density__coeff_8',
standard_deviation',
sum_of_reoccurring_data_points',
sum_of_reoccurring_values',
sum_values',
symmetry_looking__r_0.05',
symmetry_looking__r_0.1',
```

```
symmetry_looking__r_0.15000000000000002',
symmetry_looking__r_0.2',
symmetry_looking__r_0.25',
symmetry_looking__r_0.30000000000000004',
symmetry_looking__r_0.35000000000000003',
time_reversal_asymmetry_statistic__lag_1',
time_reversal_asymmetry_statistic__lag_2',
time_reversal_asymmetry_statistic__lag_3',
value_count__value_0',
value_count__value_1',
variance']
```