Modelling Complex Human Behaviours with Networks of Multi-Modal Data and Domain Knowledge available from Rich Sensor Environments

Catherine Tong

Linacre College

University of Oxford

DPhil Transfer Report

Hilary 2018-2019

Contents

Chapter 1

Introduction

1.1 Overview

Although machine learning is being widely adopted across many areas of modern technology, its success has so far been limited to a handful of well-defined domains, for instance computer vision, voice recognition and natural language processing. The vision of achieving a broad success beyond these traditional domains is currently inhibited by a number of challenges in machine learning handling *real-world* data, which may come in different forms and structures, be incomplete, time-dependent and above all messy.

The focus of this thesis is to study and push forward state-of-the-art machine learning techniques with respect to the challenges of their applications to real-world data, specifically those encountered in the health-related domain. These challenges pertain to three fundamental issues in machine learning to which this thesis is most interested in contributing: learning from graphs, learning from multiple modalities, and exploiting domain knowledge. This thesis aims to understand the current limitations and improve the designs of machine learning models in these three areas. In addition to putting forward general solutions to these challenges, this thesis will also conduct case studies focused on the application of machine learning to ubiquitous sensing for personal health and wellbeing. Through these case studies, the thesis hopes to study the practicalities of general solutions and to produce case-specific innovations.

This report is organized as follows: The remainder of Chapter 1 explains the motivations for this thesis and layout important challenges and opportunities in expanding the versatility of ML. Chapter 2 gives the research outline and plans for completed and future projects. Finally, Chapter 3 provides a literature review of the methods in machine learning in relation to the focused area.

1.2 Motivation

In recent years we have seen an unprecedented increase in our ability to generate large amounts of data, from sensors embedded in our wearable smartphone to satellite networks in space. As

the volume of data increases, so does our ability to analyze them. Machine Learning (ML), which automates the building of models on data, has become a dynamic research area with notable success in building models for data in form of text, audio and images. The influx of academic and corporate interests has rapidly transformed the field, and neural networks, known for their ability for representation learning without extensive hand-engineering and also a hunger for large data, have emerged as one of the most powerful and widely-used machine learning techniques.

A number of factors, other than the availability of big data, have contributed to the widespread success of ML applications in areas with well-defined prediction tasks. Taking image classification as an example, there are a number of benchmark datasets established for researchers to test their methods on (Deng et al. [2009], LeCun and Cortes [2010]), there is a clear underlying relationship between the image data and the prediction label, and images are conveniently structured uniformly.

Outside these well-defined domains, machine learning is a promising method that may produce a widespread positive impact. Healthcare is one such area, and one of the greatest challenges facing health systems today is chronic diseases [WHO, 2002]. Chronic diseases require a complex response from both clinicians and patients over an extended time period, but information used for patient self-management still largely comes from self-report surveys and infrequent doctor consultations. Used together with the growing volumes of health data from ubiquitous devices and electronic clinical records, ML may facilitate a wide range of health-related uses, including disease monitoring, individualized treatment recommendation, and disease prevention. In addition, since ubiquitous sensors now offer the ability for researchers to collect behavioural data in-situ, ML may also be used to analyze the collected data to further understanding of complex human behavioural patterns over time.

1.3 Challenges and Opportunities

As a data-driven discipline, machine learning relies heavily on both the quantity and quality of data it learns from. In applications of machine learning outside comfort zones of rigid data such as text and images, we see models struggling to cope with different non-uniformities of data, such as data missingness, varying modalities, individual and temporal differences and other in-the-wild uncertainties. For instance, health data may come in a variety of forms, e.g. reports, treatment procedures, images, etc. The availability of cheap and small sensors has given rise to ubiquitous sensing, where the behaviour and contexts of users may be recorded and learnt *in the wild*; in this case, data missingness might arise from individual owning different sets of sensors and using them in different times.

Apart from the ability to handle data of different qualities, machine learning, neural networks in particular, suffers from an over-reliance of large amounts of data. Since supervised learning techniques are the algorithm of choice in building many inference systems, this reliance of data means the need for large hand-labelled datasets, which is especially expensive to obtain in cases in healthcare where we need to consider domain expertise, patient privacy and burden.

While there has been a number of successful applications of machine learning for human activity and context recognition tasks using mobile phone data, two emerging trends came to our attention. First, increasing volumes of data are available from a diverse range of IoT devices from people's homes, work or other environments. Second, increasing attempts are made towards recognition tasks beyond basic recognition tasks (e.g. running) to those involving more complex behavioural phenomena, such as social interactions and disease symptoms.

How to utilize machine learning to analyze behavioural data has attracted considerable research attention in recent years; this problem is non-trivial as several challenges exist for applying traditional machine learning techniques for modelling complex human behaviours:

- **Relations and structure.** Humans are social beings, and individual actions are affected by a variety of external influences. For example, people linked together on a social network through a weight-loss app may reinforce each other's diet plan adherence. Explicitly modelling these networks, or graphs, may allow machine learning models to learn about influences between cases, and to generalize to unseen scenarios by exploiting the network topology. However, graphs lack a regular structure, making effective machine learning on graphs an open research problem.
- **Multimodalities.** Human behaviour can manifest in a great number of ways and thus behavioural data can be complicated with a diverse range of sensing modalities. For example, depression symptoms may be analyzed through patterns in sleep, communication, physical activities, amongst others. Apart from fusing together very different modalities, the varying data environments and tasks require solving problems such as selecting important information from a modality, and incorporating relationships between and within modalities.
- **Domain knowledge.** Behavioral data was traditionally analyzed by other disciplines such as psychology, medicine or social sciences. Domain knowledge can be leveraged to solve specific problems, but it is difficult to integrate structured knowledge using straightforward, hard logic rules. Many behavioural concepts may not be well-defined, there may not be a consensus in the field, or the knowledge may even be evolving with time. Defining the knowledge, transferring the knowledge, and having the model using the knowledge, all make the model design much more difficult.

Many present obstacles preventing a straightforward application is related to the fact that most current machine learning models have been developed with well-defined problems and data in mind. From the above challenges, we see that developing machine learning models which could handle data of varying qualities and quantities, with applications on non-uniform data in mind, will open the door for many techniques previously overlooked by the community.

Chapter 2

Research Proposal

2.1 Research Directions

In this thesis, we will consider fundamental issues in machine learning to make these models more applicable for widespread adoption. We break down the investigation of this thesis into three main components, in addition to case studies, as follows:

Learning from Graphs

In machine learning models of simple human activities or actions (e.g. running), the typical approach is to utilize time series information of multiple sensor data for inference. In contrast, much more comes into play for complex human phenomena, which may be influenced by important relationships not easily captured in typical data representations. In the case of an individual trying to lose weight, being connected to other people who are also trying to lose weight on social media may offer support and increase adherence to weight loss strategies, alternatively, having obese family members (who often share similar genetic background or habitats) may indicate genetic predisposition and also influence the likelihood of weight loss success. Therefore, it is important to explicitly express and model relationships amongst individuals and cases which share some form of affinity. Graphs, formed of nodes connected by edges, are simple and powerful representations which can be used to address this problem. Instead of modelling independent time series, machine learning models on graphs may be able to learn and utilize the underlying characteristics shared amongst cases based on the graph structure, to improve predictions and extrapolate for unseen scenarios.

Most deep learning models today fall short in being able to represent structure and reason about relations. In recent years, to address these concerns current research directions point towards deep learning models which are capable of learning from graphs. Within this emerging field, methods have been proposed analogous to common themes in DNNs, e.g. convolutional neural network. However, most methods have only been applied to and tested on a small number of graph datasets which are currently treated as benchmarks in the area. We observe that these benchmark graph datasets (including artificially created subgraphs of citation networks, or

graphs of molecular structure) are limited in representing the spectrum of graph characteristics, yet they are being used to validate the performance of recently proposed DNNs for graphs. The current attempts are far from considering the full set of graphs that can be encountered in the real world. For example, most benchmark graph datasets are *dense*, i.e. there exists a single largest connected component, but some graphs in ubiquitous sensing may be extremely sparse, e.g. many small disconnected components of friendships found on the social networks in apps for exercising. In this component, we focus on investigating these key questions:

1. "What are the commonly neglected or poorly-represented structural and relational information in data? If they are represented as graphs, how are the properties of the data and the model changed?"
2. "What are limitations and theoretical foundations of machine learning on a general class of graphs, which would include a spectrum of graphs with varying properties?"
3. "How can common machine learning ideas and techniques used on rigid forms of data be translated or re-invented in a graph context?"

Learning from Multi-Modalities

While many existing attempts in multimodal learning are based on traditional machine learning models (non-deep or *shallow* models), the well-known representation power of DNNs may be the answer to effective multimodal learning. Another barrier to a straight-forward implementation of current multimodal learning techniques is to do with the differences in data environments or personal habits; for instance, in health outcome predictions, data from different modalities may be available at different times for different people, and their relative interaction may also differ individually and temporally. Thus, in this component, we focus on developing multimodal learning methods which are versatile in working with vastly different modalities and data environments. These modalities may include sources of data at device or sensor level, processed or raw signals, and data environments may include settings where there are mix-sourced modalities or where modalities are absent during training or testing phase for different individuals at different times. We will address these questions:

4. "What are the capabilities of deep learning, as opposed to, shallow multimodal learning frameworks in handling a variety of multimodal data settings?"
5. "What are the limiting modality combinations or data environments where multimodal learning currently works poorly on? And how can models be improved to increase their versatility in these situations?"

Exploiting Domain Knowledge

Knowledge available from domain expert or common sense can be used not only to guide machine learning architecture design but to be explicitly encoded to facilitate machine learning. For example, the notion that a conscientious person is more likely to have a regular lifestyle could

be a useful piece of information for a model learning people's personalities given their IoT sensor data. Possible benefits of encoding knowledge in machine learning include lowering the need for large amounts of training data (as is typical for DNNs), which in turn could bring about reductions in time and computational complexities.

This component focuses on allowing machine learning models to harness knowledge from domain expert or intuition. In particular, we ask the question:

6. "How can knowledge be best represented, transferred and internalized by machine learning models, so as to benefit from domain expertise or intuition?"

Case Studies

Case studies are conducted alongside the other components to build a practical view of deploying ML on real-world data, specifically in the domain of ubiquitous sensing of personal wellbeing.

These case studies will involve applying machine learning to emerging types of ubiquitous sensing data, for complex predictive tasks relating to health outcome detection or wellbeing monitoring. The focus will be placed on gaining insights from practical issues commonly encountered in data handling and applying ML models; these insights will be part of a feedback loop to improve the development of more general issues tackled in the other components of this thesis. We also ask:

7. "What individual innovations can be made for the particular task? Is there a general class of scenarios which would also benefit from these innovations?"

2.2 Case Studies: Ubiquitous Monitoring of Wellbeing

The questions posed in Section 2.1 will be considered in a series of practical circumstances that have a significant impact on the real-world data to be encountered in this thesis. In this section, we describe case studies conducted to understand these practical circumstances. To address Question (7), in these case studies we have made individual innovations with respect to the task and will consider how to generalize these innovations through our insights and future work.

Overview

We conducted two separate case studies as pilot cases to explore the design and deployment of machine learning using data from ubiquitous sensors for health and well-being monitoring. The data is collected from a range of commercial smart devices produced by the *Withings* brand, including both IoT appliances (e.g. sleep trackers placed under the mattress) and mobile devices (e.g. smartwatch). Importantly, instead of using raw sensor signals, we consider *inferred* data which is the output of commercial-grade machine learning models for human activity recognition; for example, we use inferred sleep durations instead of raw sensor signals from the sleep tracker as inputs to machine learning models considered in these studies.

Collectively these results provide initial insights as to how to model data from mobile and appliances for use-cases in wellbeing. We believe that these investigations and results are timely given the rapid update by consumers for smart devices in the home, which will cause datasets of this type to be more readily available in the near future.

Big-Five Personality

We present a large-scale (9110-user) study of data from both mobile and networked appliances for Big-Five personality inference. Building on methods (viz. features and classifiers) previously successful for personality detection from mobile-only data, this investigation shows Big-Five personality can be detected with accuracies of 77% (similar levels as other studies) under this setting – despite the size and complexity (mix of mobile and appliance) of the dataset. This result acts as a study of techniques that are commonly utilized in the literature (e.g. random forests, support vector machines) but under a type of dataset previously never studied. Moreover, we offer ancillary results, in particular, as to behaviour and physical health features that correlate with mobile and appliance data and how inference accuracy alters as cohort scale and diversity change.

This work is currently in preparation for the International Conference on Pervasive Computing Technologies for Healthcare and is included in Appendix C. The early results from this project were presented as a poster at the ACM International Conference on Mobile Systems (MobiSys) 2018 and are included in Appendix D.

Tracking Health Outcomes of Multiple Sclerosis Patients

Multiple Sclerosis (MS) requires long-term disease management, but tracking patients through the use of clinical survey instruments is hindered by the high costs and patient burden involved. In this work, we investigate the feasibility of using data from ubiquitous sensing to predict MS patients' fatigue and health status, as measured by the Fatigue Severity Scale (FSS) and EQ-5D index. We collected data from 198 MS patients who are given connected wellness devices for over 6 months. We examine how accurately can the collected data predict reported FSS and EQ-5D scores per patient using an ensemble of regressors. In predicting for both FSS and EQ-5D, we are able to achieve errors aligning with the instrument' standard measurement error (SEM), as well as strong and significant correlations between predicted and ground true values.

We proposed and compared two adaptation methods, one based on adapted Gaussian mixture models previously proposed for speaker identification, another based on residual error correction. We show the latter, a simple adaptation method, greatly reduces prediction errors through the use of just 1 user-supplied ground truth datapoint. For FSS (SEM 0.7), the universal model predicts weekly scores with MAE 0.99, while an adapted model predicts with MAE 0.51. For EQ-5D (SEM 0.093), the universal model predicts weekly scores with MAE 0.091, while an adapted model predicts with MAE 0.052. Our study represents the first sets of results on tracking fatigue and health status of MS patients using ubiquitous sensing, which gives promising

prediction performance with errors aligns with the accepted range of error in the widely used clinically-validated questionnaires. Future extensions and potential applications of our results can positively impact MS patient disease management and support clinical research.

This work is currently submitted to and under review by the Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT) and is included in Appendix A.

Insights and Future Work

The experience of conducting these case studies has played an important part in motivating the work planned in the other components of this thesis. These case studies highlighted the difficulties in deploying machine learning outside traditional domains, especially in handling data missingness, multi-modalities, individual differences, temporal differences, and class imbalance. Also, despite being relatively large datasets within their respective domains, the small quantities of labelled data also prevent the use of innovations in deep learning, e.g. attention mechanism.

We will continue expanding the case studies with techniques developed with respect to the other components of this thesis. The first, personality, dataset suffers heavily from missing modality issues, and this may be included as a testing example for an application of utilising domain knowledge. In the MS initiative, we are also currently obtaining access to a large dataset in Oxford and the UK through our collaboration with Dr Matthew Craner at the Nuffield Department of Clinical Neuroscience. Again we plan to examine the domain knowledge incorporation problem using this dataset, and should the large dataset become available, we will also explore ways of learning a useful embedding from the large dataset to aid machine learning in the small data scenario.

2.3 Research Plan

In this section, we present our research proposal aiming to answer the questions posed in Section 2.1. We draw insights from case studies presented in Section 2.2 and plan for current and future work in the other research components.

2.3.1 Learning from Graphs

We are conducting initial investigations on the problem of learning from graphs and have not produced many concrete results thus far. On-going investigations are on two fronts: one project relates to proposing a novel framework for machine learning on graphs constructed from clinical data and targets the improvement in model predictions given by expressing data in graphs, as posted in Question (1); another project relates to Question (2) and investigates the sparsity in graphs, a neglected characteristic in most graph benchmark datasets.

Current Work

We are currently conducting two projects to investigate ML on graphs. The first is a collaboration with the University of Cambridge, focused on developing a clinical predictions framework evaluated using a publicly available clinical dataset, MIMIC-III [Johnson et al., 2016]. This dataset is a large database comprising of electronic health records (EHR) relating to patients admitted to the critical care unit at a large tertiary care hospital. The idea is to improve the robustness of prediction by developing a framework, which first utilizes a feature extraction pipeline on the EHR (e.g. recurrent neural networks for time series data, language processing for medical texts), followed by a convolutional neural network on a graph constructed using common diagnosis. We are currently in the stage of pre-processing the data and refining the methodology. We plan to submit a paper on this project to the 'Representation Learning on Graphs and Manifolds' Workshop in the International Conference on Learning Representations (ICLR) in March 2019.

Another project concerns the ability of current ML on graph methods to scale to different varieties of graphs. Delving into the construction mechanism of different benchmark graph datasets, we find that most are artificially created to overlook certain graph characteristics. In particular, we find that the sparsity of graphs, which concerns the disconnected components of graphs, may be an important overlooked factor. Most recent models for machine learning on graphs have emphasized results obtained on the 'largest-connected component', and our initial experiments on some leading models in the field find that node classification results for nodes in disconnected components perform on average at least 13% worse in accuracy compared to those in the largest-connected components. We aim to produce a set of representative graphs and compare the empirical performance of current models. We plan to submit a paper on this project to the International Joint Conference on Automated Reasoning (ICJAI) in February 2019.

Future Work

After gaining insights about the strengths and weaknesses of each model on different scenarios, we plan to propose a refined set of machine learning techniques that could cope adequately with different graph characteristics. These refined machine learning techniques may be the adoption of ideas common in machine learning proposed in the Euclidean data context and will relate to the investigations posed by Question (3). Likely directions include using transfer learning or adversarial neural networks to facilitate the effective sharing of learning process across disconnected components or multi-modalities.

2.3.2 Learning from Multi-Modalities

We have completed two separate pieces of investigations on multimodal learning. Addressing Question (4), we studied how deep learning models perform in different multimodal learning scenarios. Our MS case study presented a poor data environment for multimodal learning, and we proposed a framework to cope with the situation, addressing Question (5). This is the most mature part of the thesis so far.

Completed Work

DNNs for Multimodal Learning. We conducted a project studying the benefits of adopting deep learning algorithms for interpreting user activity and context as captured by multi-sensor systems. Specifically, in this project, we focus on four variations of deep neural networks that are based either on fully-connected Deep Neural Networks (DNNs) or Convolutional Neural Networks (CNNs). Two of these architectures follow conventional deep models by performing feature representation learning from a straghtforward concatenation of sensor types. This simple approach is contrasted with a deep model variant characterized by modality-specific partitions of the architecture to maximize intra-modality learning. Our exploration represents the first time these architectures have been evaluated for multimodal deep learning with wearable data – and for convolutional layers within this architecture, it represents a novel architecture entirely. Experiments show these generic multimodal neural network models compete well with a rich variety of conventional hand-designed shallow methods (including feature extraction and classifier construction) and task-specific modelling pipelines, across a wide range of sensor types and inference tasks (four different datasets). In addition, although the training and inference overhead of these multimodal deep approaches is in some cases appreciable, we also demonstrate the feasibility of on-device mobile and wearable execution. This study has been constructed to focus on multimodal aspects of wearable data modelling for deep learning by proving a wide range of empirical observations, which we expect to have considerable value in the community. We summarize our observations into a series of practitioner rules-of-thumb and lessons learned that can guide the usage of multimodal deep learning for activity and context detection.

This work was a collaboration with Nokia Bell Labs and University of Edinburgh, University of Cambridge. It has been accepted and presented at the IMWUT and Ubicomp'18, MobiUK'18, and is included in Appendix B.

Multimodal Learning with Missing Data. The issue of multimodal learning in poor data environments was explored in the presented case study of health monitoring in Multiple Sclerosis patients. The dataset suffers from a missing modality problem, as patients use the different sensing devices at different frequencies. Effectively combining multiple modalities for different patients also poses a problem, since different recorded aspects of their lifestyles may contribute differently to the health outcome for different patients at different times; case in point, sometimes a patient's sleep quality might indicate most about his/her fatigue, but for another patient it may be a combination of walking patterns and weight fluctuations, and at other times these correlations might be different for the same patient. Given the limited quantities of data in this dataset, we focused on shallow machine learning methods and proposed an ensemble framework for multimodal learning, which was able to cope with datapoints with different missing modalities. its detailed formulation is included in Appendix A.

Future Work

Our investigations have yet to lead to a universal multimodal fusion architecture which outperforms highly hand-engineered multimodal learning in every situation, so it is important to take

this into consideration and experiment with different formulations depending on the data and task. In the work conducted so far, we have investigated multimodal learning frameworks on 4 public datasets and 2 individual case studies, we aim to examine the spectrum of data sets and draw insights about the variety of data settings and the machine learning models that would work best for each scenario. Based on these insights, we will move on to proposing key components of a universal multimodal learning framework that would have the versatility in working well in different data settings.

2.3.3 Exploiting Domain Knowledge

This is least explored part of the thesis, we have conducted brief investigations into knowledge rule distillation as part of the personality inference case study. Our future work will seek to address Question (6).

Current work

Inspired by the rule distillation framework proposed by Hu et al. [2016] on sentence sentiment classification, we conducted initial investigations into applying a rule distillation framework for the task of personality classification, where we constructed rules such as 'a person is conscientious if the standard deviation in his/her bed-in times are below median' and relaxed the uncertainty about these rules. These rules are obeyed by a teacher network, which then distils the knowledge to a student network. In this rule example, the rationale was that an intuitive understanding of 'conscientiousness' means sticking to routines and therefore sticking to a more rigid sleep schedule. However, our implementation has yet to gain much traction so far, and the performance of rule-distilled models are not much improved.

Future work

In the future, instead of relying on a rigid rule distillation framework, we will look into developing a *knowledge* component within a deep neural network framework. A possible basis for this knowledge component may be the rule-obeying teacher network, a feature-extraction framework from unstructured texts from literature, or a structured knowledge graph. As seen in our current work, the knowledge may not be well-defined as an explicit deterministic rule or prior distribution, so we will tailor our knowledge component to consider ill-defined concepts. We will further consider ways of incorporating the knowledge and testing if the knowledge is internalized by a machine learning model. This is expected to have consequences in the model's reliance on data as a bid to counteract DNNs' data-hungry characteristic so that they may be applied to a broader variety of domains. It will also be interesting to investigate special cases where the data-driven insights are at odds with human knowledge and how the model should cope with

Chapter 3

Literature Review

3.1 Overview

This chapter presents the technological backdrop surrounding the machine learning against which this thesis is proposed. We provide a survey and discussion of the most prominent methods proposed in this thesis.

3.2 Machine Learning on Graphs

Graphs, geometric structures formed of nodes connected by edges, are common representations of data found in the real world. For example, relationships between people can be represented as social networks, trades as commerce networks, and molecules as biological networks. The complex relational and structural information that can be communicated by nodes and edges can collectively provide vital information, giving complicated structures in graphs which contain rich underlying value [Barabasi].

Utilizing machine learning for graph data analysis has attracted a lot of attention. Yet the problem of generalizing existing machine learning methods to graphs is non-trivial. Zhang et al. [2018] argued that four main challenges exist: First, graphs lie in an irregular domain so simple mathematical operations (e.g. convolution) cannot be easily defined. Second, graphs can come with diverse structures and machine learning tasks can also vary greatly, from node-focused problems (e.g. node classification, link prediction) to graph-focused problems (e.g. graph classification). The third challenge relates to the scalability of methods on large graphs. Finally, graphs are often representations of data connected with other disciplines (e.g. social sciences), thus requiring interdisciplinary collaboration.

Responding to these challenges, significant effort has been made towards this area. Earlier developments in graph ML focused on network embedding, which tries to embed nodes into a lower-dimensional vector space. A popular embedding technique is Node2vec, first proposed by Grover and Leskovec [2016] to make use of random walks on graphs; although the original algorithm is not able to incorporate node or edge features, many extensions have been made to

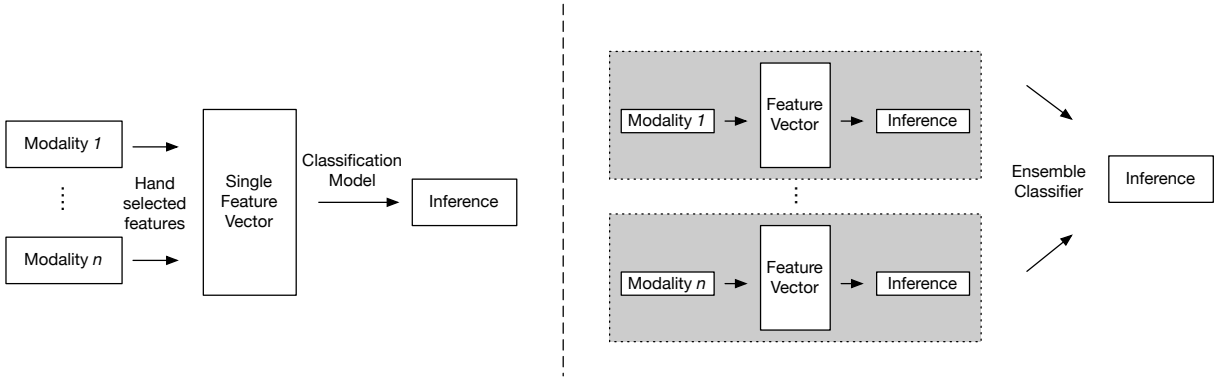
encode other desirable graph properties. In general, the inevitable shortcoming of node embedding is that some graph structure around a node may be discarded in the embedding procedure and not all node and relationship properties may be incorporated.

Recent attention in graph ML has focused on deep learning methods, especially the development of Graph Convolutional Networks (GCNs). GCNs involves storing states for each node and using an adjacency matrix to propagate those states to the neighbourhood of the nodes. Similar to Convolutional Neural Networks (CNNs), convolution is the most fundamental operations in GCNs. In the graph context, Bruna et al. [2013] pioneered convolution for graphs from the spectral domain using the graph Laplacian matrix $\boldsymbol{L}$, which plays a similar role as the Fourier basis for signal processing. This idea has since undergone multiple adaptations and improvements, notably Kipf and Welling [2016] further simplifies the convolutional filters used in GCNs by only considering the first-order neighbourhoods of nodes. Some recent work has looked into migrating other innovation from general deep learning to the graph context, including using attention [Veličković et al., 2017, Peng et al., 2018]. Some early adoption of these methods have already been made in other domains through the use of learning from knowledge graphs, for example, in health, Shang et al. [2018] uses GameNet to perform medication recommendation and integrates drug-drug interaction knowledge graphs by a memory module implemented as GCNs, and achieve good performance on large-scale EHR data. integrates

3.3 Learning from Multimodalities

Multimodal learning has a vast application domain. Applications have been seen in audio-visual speech recognition [Ngiam et al., 2011], image captioning [Sohn et al., 2014], machine translation [Kiros et al., 2014], sentiment analysis [Poria et al., 2016] and affect recognition [Kapoor and Picard, 2005]. In the space of ubiquitous computing, example applications include human activity recognition [Barz et al., 2016], sleep detection [Chen et al., 2017] and emotion recognition [Kye et al., 2017]. Many recognition tasks were previously only primarily performed with unimodal learning, with the availability of low-energy sensors, many such tasks are recently explored using multimodal learning. For example, authentication models involve both gaze and touch recognition [Khamis et al., 2016], or eating recognition might involve motion of head, wrist and audio [Merck et al., 2016].

Whilst there are numerous applications of multimodal learning, how best to combine sensor inputs which are significantly different remains an open problem. In existing works of multimodal learning models, we observe two broad categories of sensor fusion: Feature Concatenation and Ensemble Classifier. In feature concatenation, data from different modalities are simply collapsed together into single feature vectors. In ensemble classifier, modality-specific classifiers are included in an ensemble; the schematics of both frameworks are depicted in Figure 3.1. Currently, the choice of modality fusion scheme is still determined and justified on a case by case basis (e.g. Snoek et al. [2005]). Deep learning models, which are capable of feature representation learning with little hand-engineering and preprocessing, may be a promising general solution to multimodal learning.



(a) Feature Concatenation (b) Ensemble Classifier

Figure 3.1: Schematic of common approaches to shallow multimodal learning with (a) Feature Concatenation and (b) Ensemble Classifier models. For Feature Concatenation hand-selected features are extracted from each sensing modalities and concatenated into a single features vector as input to a classifier, while the Ensemble Classifier approach performs detections on each sensing modality independently to combine their estimations as final inference.

A prime example in the general space of multimodal deep learning is audio-visual speech recognition [Mroueh et al., 2015], where much work has been done using neural networks [Ngiam et al., 2011]. A number of neural networks have been proposed to perform multimodal deep learning, including CNN [Münzner et al., 2017], RBM [Ngiam et al., 2011] and RNN [Mao et al., 2014]. The choice of neural network often depends on the type of recognition involved, as there is currently no consensus on which network would work best. For instance, in tasks where sequential data is involved (e.g. image sentence description [Mao et al., 2014]), multimodal versions of recurrent neural networks have been frequently proposed to handle these tasks. While there is work comparing a small number of multimodal learning methods, such as Brown et al. [1993] which compares decision tree classifiers with backpropagation neural networks, we note that there has not been a comprehensive case study comparing a greater number of deep and shallow multimodal learning architectures.

3.4 Exploting Knowledge

Hinton et al. [2015] pioneered a technique known as 'Knowledge Distillation', which attempts to deliver smaller machine learning models by transferring the knowledge from a larger, more complex model to a smaller one. Hinton et al. argue that the soft outputs of the last softmax layer of a neural network represent 'knowledge' that the model has learnt during training, and these soft outputs can be used as a ground truth dataset for training a compact model, so that the smaller model may mimic the large one.

Whilst most work utilising knowledge distillation has been focused on the model compression area, the concept of transferring knowledge between models is useful and has inspired work that seeks to bring in human knowledge to model training. Combining deep neural networks with structured knowledge has been of increasing interests to increase generalization and improve interpretability (Deng et al. [2014]). Hu et al. [2016] proposed a framework to transfer logic

rules into neural networks with diverse architectures (such as convolutional networks and recurrent networks) using an iterative distillation framework that trains a neural network to emulate predictions of a 'teacher' model iteratively constructed by imposing posterior constraints on the network. Although this framework has been shown to be effective in regulating different neural models for incorporating grammatic logic rules for sentiment classification in text, since the method requires fixed constraints and manually specified weights, it still falls short in being able to incorporate large amounts of human intuitions or knowledge which may be ill-defined.

On a different line of work, Hong et al. [2018] has more recently proposed RDPD, a framework which utilizes knowledge distillation to tackle the missing modality problem and discrepancies between rich and poor data environments. RDPD proposes enhancing a small model trained on poor data (namely where only a single modality is available) with a complex model trained on rich data (where multimodalities are available), and has been applied to real-world datasets formed of data from the healthcare domain. Promising results have been given showing significant performance improvement after the procedure was applied.

3.5 Ubiquitous Sensing for Health Monitoring

In recent years, increasing interest has been placed on using ubiquitous sensing technology in monitoring symptoms for a number of which diseases, many of which have been carried out to reproduce predictions for clinically-validated self-report instruments in the concerned disease. For example, Wang et al. [2017] develops a prediction system that tracks schizophrenia symptoms based on a standard instrument using passive sensing from mobile phones, they were able to accurately predict reported schizophrenia scores using Gradient Boosted Regression Trees (GBRT). Wang et al. [2018] proposed a new approach in predicting depression using passive sensing data from college students' smartphone and wearables through the use of a proposed set of symptom features, they used generalized linear mixed model (GLMM) to predict self-reported depression scores and found correlations between their proposed symptom features and depression scores. Other than disease monitoring, a number of studies have also been carried looking into monitoring of more general wellness indicator, since this could be applied to the general non-clinical population, the sample dataset sizes studied could be much bigger, therefore also allowing more advanced techniques to be applied (e.g. deep neural networks). Veličković et al. [2018] analyzes multimodal time-series data and predicts the ability to achieve weight objective for users of smart connected devices using deep long-short-term memory architectures.

Besides medical disease monitoring, using ubiquitous sensing to collect behavioural data has also gained much academic attention to improve understanding in behavioural and psychological sciences. Harari et al. [2016] advocates the increased use of smartphones as a behavioural observation tool in psychological science, and outlines practical considerations and opportunities for such interdisciplinary research work. Indeed a rising number of studies have been conducted in relation to this line of work using smartphone or IoT data. Olguín et al. [2009] uses low-level sensor data from sociometric badges to study individual and group behaviour, using high-level behavioural descriptions such as physical and speech activity; their results show that it

is possible to correlate high-level behavioural information with personality traits. Staiano et al. [2012], de Montjoye et al. [2013], Chittaranjan et al. [2011, 2013] utilised data from smartphones for personalty inference, using information such as call logs, use of Bluetooth. Staiano et al. [2012] focuses on inference through building a picture of the social network of smartphone users, while de Montjoye et al. [2013] proposes behavioural indicators for inference, e.g. regularity and diversity found in calls/ texts, spatial behaviour from GPS signals.

Bibliography

A.-L. Barabasi. *Network Science*.

Michael Barz, Mohammad Mehdi Moniri, Markus Weber, and Daniel Sonntag. Multimodal multisensor activity annotation tool. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, pages 17–20, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4462-3. doi: 10.1145/2968219.2971459. URL http://doi.acm.org/10.1145/2968219.2971459.

Donald E. Brown, Vincent Corruble, and Clarence Louis Pittard. A comparison of decision tree classifiers with backpropagation neural networks for multimodal classification problems. *Pattern Recognition*, 26(6):953 – 961, 1993. ISSN 0031-3203. doi: https://doi.org/10.1016/0031-3203(93)90060-A. URL http://www.sciencedirect.com/science/article/pii/003132039390060A.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral Networks and Locally Connected Networks on Graphs. *arXiv e-prints*, art. arXiv:1312.6203, December 2013.

W. Chen, A. Sano, D. L. Martinez, S. Taylor, A. W. McHill, A. J. K. Phillips, L. Barger, E. B. Klerman, and R. W. Picard. Multimodal ambulatory sleep detection. In *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 465–468, Feb 2017. doi: 10.1109/BHI.2017.7897306.

G. Chittaranjan, J. Blom, and D. Gatica-Perez. Who's who with big-five: Analyzing and classifying personality traits with smartphones. In *2011 15th Annual International Symposium on Wearable Computers*, pages 29–36, June 2011. doi: 10.1109/ISWC.2011.29.

Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, 17(3):433–450, Mar 2013. ISSN 1617-4917. doi: 10.1007/s00779-011-0490-1. URL https://doi.org/10.1007/s00779-011-0490-1.

Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, and Alex (Sandy) Pentland. Predicting personality using novel mobile phone-based metrics. In Ariel M. Greenberg, William G. Kennedy, and Nathan D. Bos, editors, *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 48–55, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *European Conference on Computer Vision*, 2014.

Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 855–864, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939754. URL http://doi.acm.org/10.1145/2939672.2939754.

Gabriella M. Harari, Nicholas D. Lane, Rui Wang, Benjamin S. Crosier, Andrew T. Campbell, and Samuel D. Gosling. Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, 11(6):838–854, 2016. doi: 10.1177/1745691616650285. URL https://doi.org/10.1177/1745691616650285. PMID: 27899727.

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

Shenda Hong, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. RDPD: Rich Data Helps Poor Data via Imitation. *arXiv e-prints*, art. arXiv:1809.01921, September 2018.

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. Harnessing Deep Neural Networks with Logic Rules. *arXiv e-prints*, art. arXiv:1603.06318, March 2016.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035 EP –, May 2016. URL https://doi.org/10.1038/sdata.2016.35. Data Descriptor.

Ashish Kapoor and Rosalind W Picard. Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 677–682. ACM, 2005.

Mohamed Khamis, Florian Alt, Mariam Hassib, Emanuel von Zezschwitz, Regina Hasholzner, and Andreas Bulling. Gazetouchpass: Multimodal authentication using gaze and touch on mobile devices. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 2156–2164, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4082-3. doi: 10.1145/2851581.2892314. URL http://doi.acm.org/10.1145/2851581.2892314.

Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv e-prints*, art. arXiv:1609.02907, September 2016.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014. URL http://arxiv.org/abs/1411.2539.

Saewon Kye, Junhyung Moon, Juneil Lee, Inho Choi, Dongmi Cheon, and Kyoungwoo Lee. Multimodal data collection framework for mental stress monitoring. In *Proceedings of the 2017*

ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers, UbiComp '17, pages 822–829, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5190-4. doi: 10.1145/3123024.3125616. URL http://doi.acm.org/10.1145/3123024.3125616.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain Images with Multimodal Recurrent Neural Networks. *ArXiv e-prints*, October 2014.

Christopher Merck, Christina Maher, Mark Mirtchouk, Min Zheng, Yuxiao Huang, and Samantha Kleinberg. Multimodality sensing for eating recognition. In *Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '16, pages 130–137, ICST, Brussels, Belgium, Belgium, 2016. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). ISBN 978-1-63190-051-8. URL http://dl.acm.org/citation.cfm?id=3021319.3021339.

Y. Mroueh, E. Marcheret, and V. Goel. Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2130–2134, April 2015. doi: 10.1109/ICASSP.2015.7178347.

Sebastian Münzner, Philip Schmidt, Attila Reiss, Michael Hanselmann, Rainer Stiefelhagen, and Robert Dürichen. Cnn-based sensor fusion techniques for multimodal human activity recognition. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, ISWC '17, pages 158–165, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5188-1. doi: 10.1145/3123021.3123046. URL http://doi.acm.org/10.1145/3123021.3123046.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 689–696. Omnipress, 2011.

Daniel Olguín Olguín, Peter A. Gloor, and Alex Pentland. Capturing individual and group behavior with wearable sensors. In *AAAI Spring Symposium: Human Behavior Modeling*, 2009.

Hao Peng, Jianxin Li, Qiran Gong, Yuanxing Ning, and Lihong Wang. Graph Convolutional Neural Networks via Motif-based Attention. *arXiv e-prints*, art. arXiv:1811.08270, November 2018.

Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174(Part A):50 – 59, 2016. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2015.01.095. URL http://www.sciencedirect.com/science/article/pii/S0925231215011297.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. GAMENet: Graph

Augmented MEmory Networks for Recommending Medication Combination. *arXiv e-prints*, art. arXiv:1809.01852, September 2018.

Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM, 2005.

Kihyuk Sohn, Wenling Shang, and Honglak Lee. Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2141–2149, 2014. URL http://papers.nips.cc/paper/5279-improved-multimodal-deep-learning-with-variation-of-information.

Jacopo Staiano, Bruno Lepri, Nadav Aharony, Fabio Pianesi, Nicu Sebe, and Alex Pentland. Friends don't lie: Inferring personality traits from social network structure. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 321–330, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1224-0. doi: 10.1145/2370216.2370266. URL http://doi.acm.org/10.1145/2370216.2370266.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *arXiv e-prints*, art. arXiv:1710.10903, October 2017.

Petar Veličković, Laurynas Karazija, Nicholas D. Lane, Sourav Bhattacharya, Edgar Liberis, Pietro Liò, Angela Chieh, Otmane Bellahsen, and Matthieu Vegreville. Cross-modal recurrent models for weight objective prediction from multimodal time-series data. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '18, pages 178–186, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-6450-8. doi: 10.1145/3240925.3240937. URL http://doi.acm.org/10.1145/3240925.3240937.

Rui Wang, Weichen Wang, Min S. H. Aung, Dror Ben-Zeev, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A. Scherer, and Megan Walsh. Predicting symptom trajectories of schizophrenia using mobile sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):110:1–110:24, September 2017. ISSN 2474-9567. doi: 10.1145/3130976. URL http://doi.acm.org/10.1145/3130976.

Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1):43:1–43:26, March 2018. ISSN 2474-9567. doi: 10.1145/3191775. URL http://doi.acm.org/10.1145/3191775.

WHO. Innovative Care for Chronic Conditions: Building Blocks for Action. 2002.

Ziwei Zhang, Peng Cui, and Wenwu Zhu. Deep Learning on Graphs: A Survey. *arXiv e-prints*, art. arXiv:1812.04202, December 2018.

Appendix A

Conference Paper: Ubicomp 2019

This work was co-supervised by Nic Lane and Matthew Craner, a neurologist based at the Nuffield Department of Clinical Neuroscience, University of Oxford. We have recently submitted this work to the November cycle of the Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT).

Tracking Fatigue and Health State in Multiple Sclerosis Patients Using Ubiquitous Sensing

Multiple Sclerosis requires long-term disease management, but tracking patients through the use of clinical survey instruments is hindered by the high costs and patient burden involved. In this work, we investigate the feasibility of using data from ubiquitous sensing to predict MS patients' fatigue and health status, as measured by the *Fatigue Severity Scale (FSS)* and *EQ-5D* index. We collected data from 198 MS patients who are given connected wellness devices for over 6 months. We examine how accurately can the collected data predict reported FSS and EQ-5D scores per patient using an ensemble of regressors. In predicting for both FSS and EQ-5D, we are able to achieve errors aligning with the instrument' standard measurement error (SEM), as well as strong and significant correlations between predicted and ground true values. We also show a simple adaptation method that greatly reduces prediction errors through the use of just 1 user-supplied ground truth datapoint. For FSS (SEM 0.7), the universal model predicts weekly scores with MAE 0.99, while an adapted model predicts with MAE 0.51. For EQ-5D (SEM 0.093), the universal model predicts weekly scores with MAE 0.091, while an adapted model predicts with MAE 0.052. Our study represents the first sets of results on tracking fatigue and health status of MS patients using ubiquitous sensing, which gives promising prediction performance with errors aligns with the accepted range of error in the widely used clinically-validated questionnaires. Future extensions and potential applications of our results can positively impact MS patient disease management and support clinical research.

ACM Reference format:
. 2018. Tracking Fatigue and Health State in Multiple Sclerosis Patients Using Ubiquitous Sensing. 1, 1, Article 1 (November 2018), 24 pages.
https://doi.org/0000001.0000001

1 INTRODUCTION

Over 2.5 million people worldwide are affected by Multiple Sclerosis (MS), in many countries it is the most common cause of neurological disability in young adults [28]. While not fatal, Multiple Sclerosis is a chronic debilitation disease associated with significant health-related and economic burden to the quality of life [16]. MS has also been described as a highly individual disease, characterized by a variety of disabling symptoms experienced by different patients, although excessive fatigue is the most frequent symptom [11]. As there is currently no cure for MS, it is a life-long disease that affects the everyday lives of patients which they have to learn to cope with. The complexity of MS demands that patients be active in their management of symptoms and receive support for it [10].

Managing MS mainly involves keeping track of clinical outcomes, such as symptoms and health-related quality of life. One of the most commonly reported symptoms of MS is overwhelming *fatigue*, and it is also amongst the most frequently monitored. MS Fatigue is very different from what persons with MS may normally experience after exercise, and often has a profound impact on patients' quality of life [39]. Keeping a longitudinal record of fatigue is not only important for patients' self-management of the symptom, but also for developing treatment methods and futhuring understanding of the disease. The Fatigue Severity Scale (FSS) is the most commonly used unidimensional scale for measuring fatigue, it is a widely clinically validated questionnaire used by millions

XXXX-XXXX/2018/11-ART1
https://doi.org/0000001.0000001

of patients. In clinical trials, patients' fatigue is reported longitudinally using FSS to demonstrate the efficacy of drugs [17]. A large body of research also exists to address correlations between fatigue and other physiological factors (e.g. structural damage to the brain [4]) and behavioral factors (e.g exercise, temperature [6, 39]). However, since data collection is costly and slow, in longitudinal studies of MS fatigue, the data may only be collected annually for 2-3 years. Many studies have noted the scarcity of more fine-grained records, pointing out the limitations in sparse observations, which might lead to overlooking of important dynamic fluctuations [29].

For medical and economic researchers, an important measurement to collect is the *Health-Related Quality of Life (HRQoL)*. HRQoL is typically measured through self-reported questionnaires by the EQ-5D, which can be interpreted as the *health state* pf a person. Policymakers make use of HRQoL to evaluate the effectiveness of different health intereventions, so that resources are allocated effectively [8]. For MS, this means allocating more resources to treatment that leads to a delay in patients' progression to permanent disability, over treatment which only temporarily avoids disability [16]. Longitudinal tracking of this metric is consequential to the resource allocation decision making, however frequent data collection is again an issue (data collected over clinical trials are deemed too short) so researchers typically rely on modelling for economic evaluations in multiple sclerosis [16].

It is important for MS patients and researchers to track *fatigue* and *health state* over time, and researchers are looking to track such metrics over longer terms and at more fine-grained intervals but are having difficulties due to the high costs associated with data collection. Given this, we propose a study looking into tracking fatigue and health state unobtrusively in the daily lives of MS patients through the use of connected wellbeing devices. We are interested in applying machine learning techniques in delivering accurate predictions of pateints' fatigue and health state over time in a low user burden fashion, which could support patient's self-management and further understanding of more dynamic patterns of these health indicators.

We conducted a 6-month study of 198 MS patients which collected their behavioral, physiological and self-reported health information. Each patient was given 3 connected wellness devices (in-bed sleep tracker, smart scale, smart watch) to use in their natural environments, these collect daily behavioral and physiological data of the patient. In addition, through a companion app, patients may fill in daily evaluations of fatigue and sleep quality, as well as weekly questionnaires containing the FSS and EQ-5D. In this work, we examing predicting FSS and EQ-5D scores and in particular restrict ourselves to 2 tasks per index: 1. predicting the mean score over all weeks reported by each patient, 2. predicting the score reported each week by each patient. We propose using an ensemble of device-specific predictors for both tasks to improve performance and to handle issues arising from missing modalities. To better account for person-to-person variations and improve generalization performance, we further propose ways of personalization of the ensemble model, through the use of a minimal number of self-reported scores by each patient.

The main contributions of this work is as follows:

- First study predicting MS patients' reported fatigue severity scale (FSS) using data from connected devices and daily reports at weekly intervals. We show that FSS (instrument error SEM 0.7) can be predicted with MAE 0.99 with a universal ensemble model, and with MAE 0.51 with an adapted model. These results are promising as they are within scales of the instrument error and also correlate significantly and strongly with ground truths.
- First study predicting MS patients' reported EQ-5D health state using data from connected devices and daily reports at weekly intervals. We show that EQ-5D (SEM 0.093) can be predicted with MAE 0.91 with a universal ensemble model, and with MAE 0.052 with an adapted model. These results are promising as they are within scales of the instrument error and also correlate significantly and strongly with ground truths.
- Examination of data adaptation methods under this sparse data scenario. We compare a novel application of the MAP-adapted Gaussian Mixture Regression Models with a simple residual error translation. We

show that the latter (using just 1 week of user's reported ground truth) is a pragmatic approach given the high cost of data collection and it significantly improves model performance by 51% on average for FSS and EQ-5D predictions.
- Investigation of predicting the per-participant averaged scores for FSS and EQ-5D over the 6-month study period, which represents a more stable view. We show that the mean FSS can be predicted with MAE 0.95, and mean EQ-5D can be predicted with MAE 0.84.
- Discussion and summary of insights from current study for future work in related directions.

2 BACKGROUND

In this section, we provide some background information about MS. Multiple Sclerosis is a chronic, often disabling, disease of the Central Nervous System (CNS). As MS causes damage in the CNS, it can adversely affect nearly any body function, the most commonly experienced symptoms include overwhelming fatigue, visual disturbances, altered sensation and difficulties with mobility. From a clinical perspective, the course of MS is highly varied and diverse, varying in types and severity between persons and within a person over time. Persons with MS have a prolonged median survival time from diagnosis of around 40 years, and over this time the presentation of MS changes. Most patients are initially diagnosed with Relapsing-remitting MS (RRMS), where they experience relapses separated with periods of remissions. Many patients, over a span of 5 to 15 years, gradually transform into steady deterioration, known as the Secondary progressive MS (SPMS). Two other rarer forms of disease course are Primary-progressive MS (PPMS) and Progressive-Relapsing Multiple Sclerosis (PRMS). [10, 28]

Treatment. As of yet, there is no cure for MS. Modern treatment focuses on the management of symptoms and disease course, but are only partially effective. To relieve symptoms and to promote a satisfactory quality of life, MS requires patients to be active in the management of their own health [10]. Self-managing typically involves keeping a symptom diary, where patients manually record their symptoms or activities to keep track of their disease course, and there are a number of mobile apps developed which incorporate such diary function. However, monitoring is a tedious process and as a reuslt can yield low patient adherence; in addition, it is prone to errors due to the subjective nature of symptom recall [7], compounded with the possibility that some MS patients might also experience cognitive symptoms such as memory difficulties.

Measurements. It is important to measure both fatigue and quality of life with robust, clinically validated instruments as these are subjective phenomena which can only be measured subjectively. When using these standard survey instruments, the health community also looks at their associated standard error of measurement (SEM) to access the reliability of the instrument. FSS and EQ-5D both widely used instruments in the MS community, with studies reporting SEMs of 0.7 and 0.093 respectively [20, 24].

Measuring Fatigue. A frequently used method for evaluating fatigue is the 9-item Fatigue Severity Scale (FSS) developed by Krupp et al for use in patients with MS and systemic lupus [18], but has since been used frequently in different populations of people with chronic illnesses [2, 14]. Fatigue Severity Scale (FSS) asks the patient to rate their fatigue according to experiences of the previous week. Such clinically validated instruments are typically used in doctor-patient interactions for doctors to benchmark changes in individuals during check-ups at outpatient clinics, so this could mean FSS is typically collected about bi-annually despite being a weekly instrument. However, in patient's personal records of fatigue, such as their fatigue diary, patients typically use arbitrary metrics that they find the easiest to record in and this often means simply a numerical rating of fatigue. Efforts have been put into developing validated and quick measurements of fatigue in real-time, [15] asked 49 MS patients to wear a wrist-worn device preprogrammed to beep 4 times a day so that the patient would enter his/her fatigue level (from 0 -10) into the device real-time. However other metrics have yet to gain traction and FSS remains to most widely used metric in MS studies.

Measuring Quality of Life. The EQ-5D assessment instrument is widely used for measuring health-related quality of life (HRQoL). EQ-5D covers 5 dimensions of health: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression [34]. There are two levels, 3-level EQ-5D (EQ-5D-3L) and 5-level 5Q-5D (EQ-5D-5L), where level refers to the length of the Likert-Scale. The EQ-5D-5L, developed by Euroqol group, was shown to have better discriminative capacity and sensitivity to changes in health status, as well as smaller ceiling effect than the original 3-level EQ-5D [8]. It is a widely validated and assessed metric used in studies of MS and other diseases. Country-specific value sets, reflecting trade-offs that individuals are willing to make between health outcomes, were used to convert responses on the 5Q-5D-5L to a single-value health state index. This index is anchored at 1 (perfect health) and 0 (death). Valuations less than zero reflecting health states âĂŸworse than deathâĂŹ (WTD), can exist.

3 ANTICIPATED USAGE SCENARIOS FOR AUTOMATED MS TRACKING

Technology can support MS patients by offering useful tools to monitor their symptoms. In particular, the proliferation of ubiquitous computing devices enables continuous personal monitoring, which may be extrapolated for personal health management. This is part of the paradigm of Connected Health, where new models of health management are rising to let the patient become the centre of the health care system, supported by ubiquitous sensing.

Obtaining such longitudinal, granular, and objective records of patient conditions is valuable to both patients and health care professionals. For patients, this information can supplement their current records, alleviates their burden in record-keeping, and more importantly, they may use this to better understand their conditions, identify symptom triggers leading to activity modification and lifestyle changes, and experiments with self-management strategies with objective feedback. Case in point, most MS patients are physically inactive despite evidence of exercise training being beneficial to their disease[19], a possible scenario is patients could use these fatigue patterns to identify a good time for exercising, and to review how their fatigue levels change following different exercise sessions. For health care professionals, they could work with patients using such more thorough history to suggest activity modification and lifestyle changes that optimize their overall quality of life. When prescribing treatment, they may also use this information to judge the effectiveness and suitability of treatment for the individual patient.

On a research level, this represents a new stream of information that could allow researchers to better understand the disease course with data from more patients and at finer timescales. This could bring potential benefits in a number of areas in MS which has yet to be elucidated, e.g. the transformation from RRMS to SPMS in an MS patient is poorly understood but has important implications for treatment as many drugs effective when targeted at RRMS seem useless with SPMS [28].

Whilst this study remains the first step towards a fully-deployable system available for MS patients, we envision that such a system would have the following features:

- *Universal model for tracking MS symptoms with the ability to personalize.* When users first started using the smart devices, predictions would be made using a universal model. Over time the model would better adapt to the user's data based on accumulation of data.
- *Continuous analysis.* Behavioral and physiological passive sensing data from user's smart devices would be continuously collected, and automatically uploaded to a secure server for processing or processed locally on the device. Weekly feedback reports could be generated and available for the user to view in a companion phone app.
- *Remote monitoring.* Data and predictions visualizations for health care professionals or other parties that the patient might choose to share this information with.

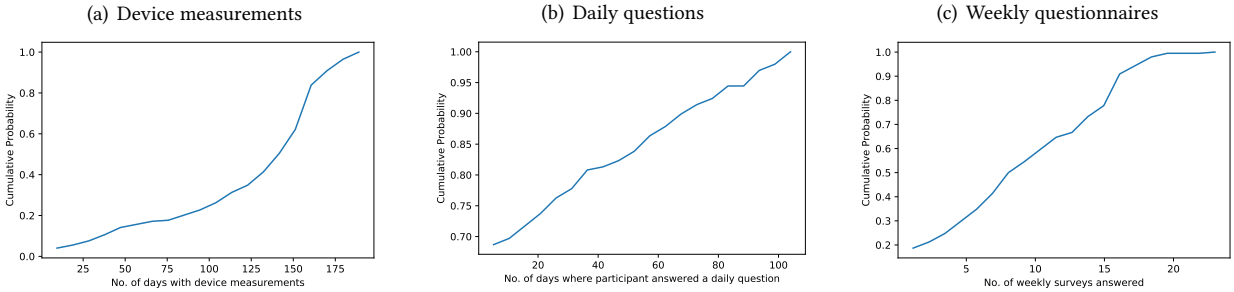


Fig. 1. CDFs of amount of data collected over the population of 198 participants.

- *Suggesting personalized recommendations.* Based on device measurements, FSS predictions, and availability of other information, e.g. weather, suggest possible correlations and triggers for patients (and their doctors) to review.

4 DATA COLLECTION AND PROCESSING

In this section, we describe the study carried out, the data collected and preprocessing done.

4.1 MS Study

We conducted a study with 198 patients diagnosed with Multiple Sclerosis who are based in the United States. The participants were recruited online for a 6-month study which seeks to understand their MS condition using connected health devices, which the participant could keep after study completion. The devices are commercial products under the *Withings* range, namely *Aura* (sleep tracker), Activité Steel (smart watch) and Body Cardio (smart scale). At the beginning of the study, each participant was given the three devices and was asked to complete a screener survey on background information about their MS. Participants were informed that the measurements on their devices over the study duration would be collected and analysed, they are free to deploy the devices at home (or outdoors in case of smart watch) and use them as frequently or infrequently as they would like. Through a companion app, patients may review their data, and also be prompted to answer questionnaires about their condition periodically. A weekly questionnaire contains two clinically valid questionnaires widely used in understanding MS (Fatigue Severity Scale (FSS) and the EQ-5D-5L), whereas two daily questionnaires, with one question each, ask the participant to rate his/her fatigue and sleep quality.

4.2 Data

The data studied consists of the following components: 1. Daily device measurements, 2. Daily reports, 3. Screener questionnaire, 4. Weekly questionnaires. For the purposes of this work (i.e. predicting FSS and EQ-5D scores), we build a universal model for the entire pool of participants using data from the first 3 components, and obtain the ground truths for these scores from the completed weekly questionnaires of each participant. Figure 1 shows the cumulative distributions of the amount of data collected for each of the components (except for the screener questionnaire which every participant has to complete).

Daily device measurements. Three products from the *Withings* range [1] (sleep tracker, smart watch, smart scale) provide measurements that may capture the time-varying quality of the MS condition relating to sleep, activity, and physiological changes. These devices deliver a 'new type of data', referring to the derived or inferred

Fig. 2. Device data is collected from watch, aura, and scale. Daily and weekly surveys are collected through the app.

data produced using machine learning algorithms associated with each device. We analyze this derived data, consisting of physiological and behavioral measurements aggregated over each day (e.g. step count of a day) so as to increase the explainability of this work given its medical context. An exhaustive list of the measurements produced by each device and associated modalities is provided in Table 1. For each participant, 189 days of their device data were extracted for analysis from November 2016 to May 2017. Device usage varies, with the participants contributing a mean of 122.8 days with at least one valid device measurement; there are 3 participants who produced no valid device data at all during the study period.

- *Withings Aura (sleep tracker).* The Aura system consists of two devices, a sleep sensory pad placed underneath the mattress, and a bedside device with environmental sensors which can act also as a smart alarm clock, speaker and lamp. Aura provides daily inference data on sleep quality and heart rate.
- *Withings Activité Steel (smart watch).* The Activité Steel is a wearable device which tracks activities (gait, running, swimming and sleeping). We only consider data on gait and sleeping as proxies to participants' activities and sleep respectively.
- *WIthings Body Cardio (smart scale).* This scale provides readings on weight, fat, muscle, water and heart health. The scale provides pulse wave velocity (PWV), the rate at which blood pressure pulse propagates through the circulatory system, which is an important clinical parameter for evaluating cardiovascular risk [25]. The scale can be set up for multiple users in case the participant's family might use it, so that data for each user stays separate, however only data from the patient's profile were collected.
- *Health Mate App.* This is a commercial app developed for use in conjunction with *Withings* devices currently used by millions of users. With the app, users create a profile to which data gathered from all *Withings* products will be synced. The user may use the app to visualize and review the data, set goals and reminders, or manage their devices [23]. Although not a focus of the current study, additional functions available on the app include leaderboard and reward badges for certain wellness achievements (e.g. step count). In this study, the app is also a portal for participants to access the daily and weekly questionnaires.

Note on Modalities. We group measurements into modalities shown in Table 1 which are meaningful together (e.g. 'aura' for sleep behavior collected). This is also to account for the gaps in our dataset, so that we separate measurements which do not always appear together into different modalities, this is seen in measurements of body composition, where a valid fat mass may not always be available. Also, we keep sleep measurements made by Aura and the watch separate, as we observed that a model trained from a mixed dataset gives poorer performance. This is in accordance with the procedure taken in similar situations [38] and relates to the heterogeneous sensing capabilities and machine learning algorithms associated with each device.

Daily questions. The patient may indicate his/her daily fatigue level and sleep quality on a 5-point Likert scale, from No to Extreme Fatigue, and Very Good to Very Bad Sleep. Each question is prompted independently

Table 1. Sources of data in this study, their associated modalities and measurements. Note that the sources 'Fatigue', 'Sleep', 'Weekly Survey' are actually obtained through the app but for purpose of later analysis we consider these individually.

Source	Modality	Measurements
Aura (sleep tracker)	Aura	sleep duration, bed-in/out times, time to sleep, time to wake, no. of times of being awake, duration of awake/ light/ REM/ deep sleep
	Night	night heart rate, night respiratory rate
Watch	Watch	sleep duration, bed-in/out times, time to sleep, time to wake, no. of times of being awake, duration of awake/ light/ deep sleep
	Steps	step count
	Walk	walking speed
Scale	Weight	weight
	Composition	bone mass, hydration mass, muscle mass
	Fat	fat mass
	Pwv	pulse wave velocity (pwv)
	Standing HR	standing heart rate
Faituge	Fatigue Level	daily fatigue level score (non-clinical)
Sleep	Sleep Quality	daily sleep quality score (non-clinical)
Weekly Survey	Weekly Survey	(response *usage*)
Screener Survey	Static	age, gender, height, disability level, type of ms, years since diagnosed, symptoms (e.g. fatigue, memory loss, motor deficits) other diseases (e.g. depression, heart disease, epilepsy) current and past treatment (e.g. use of disease-modifying agent, drug name) ideas about MS (e.g. understanding of the implications of gaps in treatment) technology use (usage and ownership of tech products) source of treatment finance (insurance or own pocket)

through the app, so the participant can answer only one or both of these questions. For this reason, we treat data from each question as a separate modality. In total, there are 3316 daily reports collected, however, there is low participation as more than half the participants (125) did not attempt this at all.

Screener questionnaire. This collects self-reported socio-demographic variables and clinical variables from each participant before the commencement of the study. All of the collected variables are categorical with the exception of age, height and number of years with MS. All 198 participants completed the screener questionnaire.

- *Socio-demographic variables.* This collects data on demographics, the patients' understanding of MS, use of technology to monitor health, as well as the patient's mode of financing treatment. There is a gender imbalance in the dataset (184 female, 14 male), as seen in most MS studies due to the higher risk of the

female population being affected. 184 participants reported they own a smartphone, while the remaining 4 own at least a tablet or Apple iPod touch.

- *Clinical variables.* This collects data on the patient's MS, symptoms experienced, other co-existing diseases, and their treatment method. Variables from symptoms and co-existing diseases are dichotomized into yes or no, so only a binary indication of whether symptoms/disaese are present is available. The majority of the participants (191 out of 198) suffer from RRMS, the most popular form of MS.

Weekly questionnaires (Fatigue and Health State Ground Truth). The weekly questionnaire is formed of two standard questionnaires which, in addition to being used for studying patients with a variety of disorders, have been widely used for MS study and management. Participants can access this questionnaire through the Health Mate app. A complete response to every question on the questionnaire (from both instruments) is required to qualify as a valid response. We use the FFS and EQ-5D-5L index scores as our ground truth. We use the notation FSS-W and EQ-W to refer to the FSS score and EQ-5D-5L index value obtained from each weekly survey respectively; and we use FSS-M and EQ-M to refer to the mean per-participant score of FSS-Ws and EQ-Ws obtained from all weekly surveys completed by each participant. In total, 1693 weekly surveys were completed, however 24 participants did not complete any weekly surveys at all.

- *Fatigue Severity Scale (FSS).* The FSS is a 9-item questionnaire each item is scored on a 7-point Likert scale, ranging from 'strongly disagree' to 'strongly agree'. The items relate to the severity of fatigue (e.g. My fatigue prevents sustained physical functioning.)and its effects on a person's activities and lifestyle (e.g. Fatigue interferes with carrying out certain duties and responsibilities.) It also covers the emotional implications of fatigue (e.g. My motivation is lower when I am fatigued.). The responses are rated on a 7-point Likert scale, and the overall FSS score is taken as the mean of the 9 responses. FSS score ranges from 1 to 7, where a score of 7 would mean strongly agreeing with all items, and a score of 1 strongly disagreeing with all items. A higher FSS suggests higher levels of fatigue. In the medical literature, a number of studies have been carried out validating the instrument, one study suggests that the standard error of measurement (SEM) of FSS was 0.7 points, and that a change in FSS of less than 0.7 points may be due to measurement error [20].
- *EQ-5D-5L index value.* The EQ-5D-5L index is a single index value computed from the EQ-5D-5L descriptive system, which describes a person's health state relative to others in the country. The conversion from 5-dimensional 5Q-5D-5L health state to the single value is done using publicly available country-specific value sets, and in this case, US-specific [33]. The 5Q-5D-5L index value ranges from 0 to 1, with 0 meaning death and 1 complete health [26].

4.3 Preprocessing into Features

We construct features for prediction by considering the relevant time series, and aggregating the measurements over the time series to produce useful statistics. We also compute a *usage* feature for the 6 sources (aura, watch, scale, fatigue question, sleep quality question, weekly survey) after realising that the missingness in data might be informative about participant's behavior as well.

FSS-M and EQ-M. For predicting FSS-M and EQ-M, the relevant time series is the entire 6-month time series for each participant. We aggregate measurements over the time series for mean, maximum, minimum, standard deviation, range, and *usage*. For each smart device, we compute *usage* as the proportion of days that a device is used. For the daily questions, usage is the proportion of days that the participant responded to each question. Finally for the weekly surveys, usage is the proportion of weeks that the participant has completed a weekly survey.

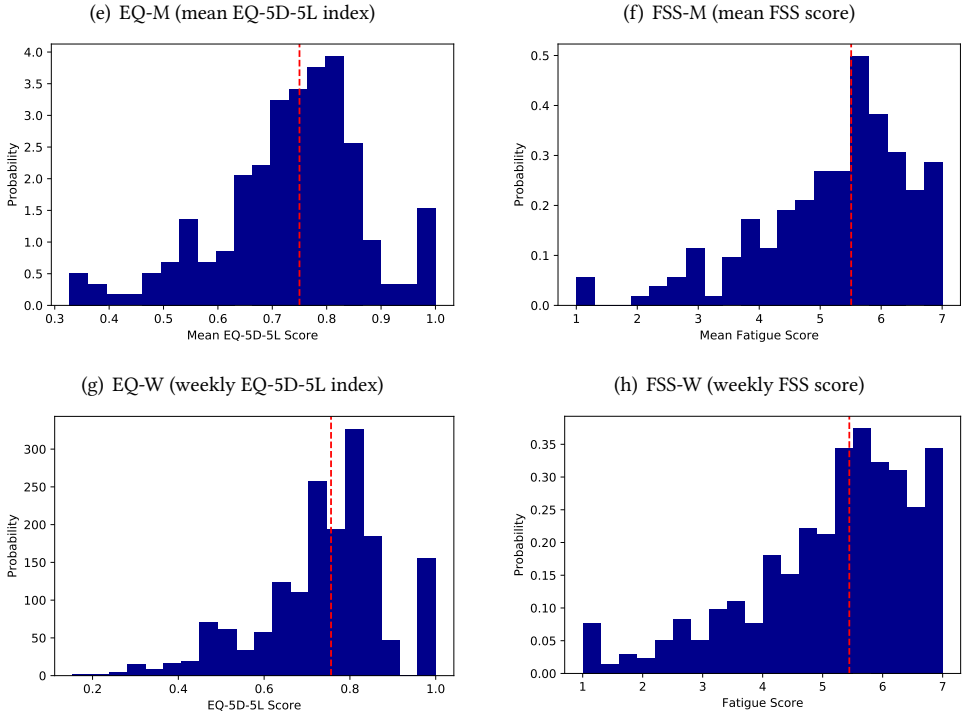


Fig. 3. Distribution of computed FSS and EQ-5DL index scores, with a red line representing the median.

FSS-W and EQ-W. For predicting FSS-W and EQ-W, the relevant time series are time windows close to the date of completing the weekly survey. For example, for a weekly survey completed on day d, we consider 2 time windows with length l before d, such that the i^{th} window spans from day $d - il$ to day $d + il$. We aggregate measurements over each time window for mean, maximum, minimum, standard deviation, range, and usage. Here, *usage* is defined slightly differently as we would like to compare the usage behavior within the relevant time window and outside the window. *Usage* is the ratio between the number of days a measurement (device/ daily question) is obtained within the window and outside the window. In addition, we also compute *usage* for the weekly questionnaires itself, defining it as the ratio between the number of completed weekly surveys so far (up till the current time window) and the number of weeks elapsed since the participant's first completed weekly survey.

Usage Features. While the missingness in data could create problems in limiting the number of datapoints, we also tried to make use of the missingness in data to represent the device usage behavior of the patients. We propose that the usage behavior of different devices or modalities could be related to the psychological decisions that the user makes and thus gives some hints about their emotional state as well. This is especially true for the usage of daily questions, as it relates to whether the user chooses to report his/her health state for the day.

4.4 Correlations of Features with Ground Truth

In Figure 4 (a) and (b), we present the correlation matrices of the mean measured features computed and FSS-W and EQ-W scores. Correlations with the computed score are presented on the top row whereas the rest are

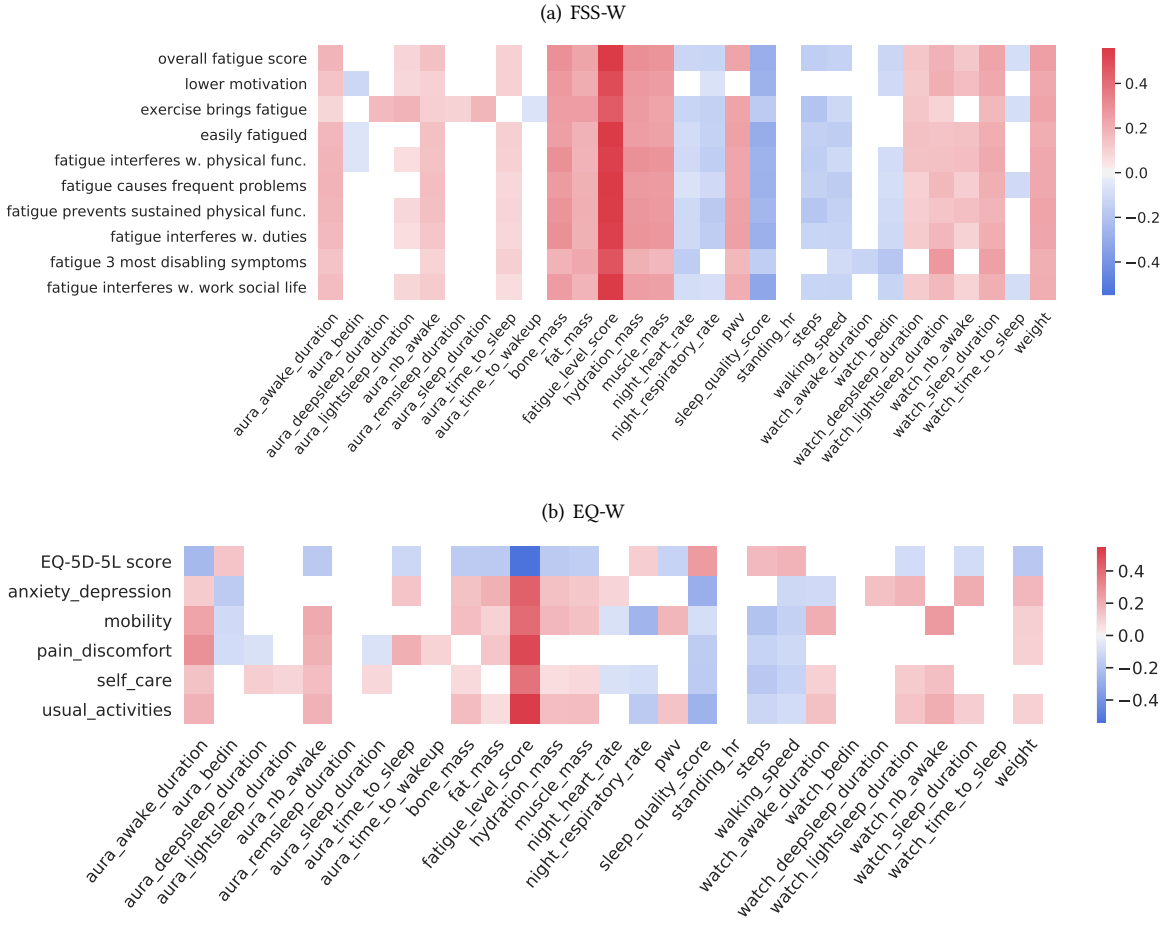


Fig. 4. This shows correlation of the mean measurements with FSS-W, EQ-W and their components. Note that the number of datapoints used to compute each correlation varies as the features are present in varying abundancies.

correlations with the sub-items scores. When viewing row by row, one can see that different dimensions of fatigue or health state are more strongly correlated with the features available. 'Exercise brings fatigue' for FSS-W and 'mobility' for EQ-W are the sub-items which has the largest average correlation with the mean features, which aligns with the common intuition that about the spectrum of wellness that such health devices may be able to capture. The correlation matrices also show that each data source has features which demonstrate significant correlations with fatigue and health states, and that the daily question of fatigue level score, despite its limited data, demonstrates a strong correlation with both FSS-W and EQ-W.

There are cases where we observe that the correlation polarity found between the same feature measured by Aura and the smart watch are opposite, namely 'time to sleep' and 'awake duration' with FSS-W. In both cases, the polarity given by the mean Aura features are overwhelmingly positive, but negative for the Watch features. One might imagine that the model computing 'time to sleep'/'awake duration' taking data from a wearable watch

being quite different from one taking data from a stationary sensor pad, especially during these stages where the user is not (yet) sleeping and is possibly moving about a lot more than during sleep. However, this ascertains our treatment of the measurements generated by the two devices as separate sources, and also raises precautions over our interpretation of any found correlations between features and ground truth, that it is best to reiterate that derived data has been used in the process.

In addition, to validate the inclusion of the *usage* feature, we also looked into the correlations between the usage features and FSS-M and EQ-M scores. In particular, we observe that the highest correlation is found at r=0.13 between the usage of the fatigue daily question and an item in the FSS-M relating to fatigue leading to lower motivation in MS patients. This aligns with our conjecture that these usage features might be informative about a user's emotional state.

5 METHODS

As an overview, we perform three main arms of investigations:

(1) *Predicting per-participant FSS-M and EQ-M.* This relates to a stable 6-month view of fatigue and health states. We compare model performance.
(2) *Predicting weekly FSS-W and EQ-W.* This relates to a dynamic weekly view of fatigue and health states. We compare performance between universal models here.
(3) *Achieving personalization via adaptation to individual user's data.* This is for predicting FSS-W and EQ-W. We compare performance between adapted models here using small amounts of user-supplied ground truth labels.

In the following, we describe our model formulations, fine-tuning and evaluation procedures for these prediction tasks.

5.1 Ensemble Model

Our main results are derived through use of an ensemble model of regressors. We explain this choice and its construction in this subsection.

As features from multiple modalities are present to a varying degree in our dataset, it is desirable to learn accurate models which can leverage information from multiple modalities while considering as many datapoints as possible. Motivated by this, we adopt an ensemble learning structure, formed by using predictors focusing on different spectrums of the data which has high data concentration. This in effect means having an ensemble of predictors trained specifically with data of the same modality, or source. Ensemble methods often offer better inference performance than that produced by its components. Unless otherwise specified, we consider an ensemble of source-specific classifiers, which use data from one of 6 sources, namely aura, scale, watch, daily fatigue question, daily sleep-quality question and static (screener survey). However in order to be considered as a valid datapoint, features must come from at least one other source besides than static, this is to ensure that time-varying data is included.

Weighting function. In combining the predictions given by each modality, we put to use the resulting feature importance vector generated per use in the feature selection phase. We compute the mean feature importance per source, arguing that this would give a good representation of prediction confidence of each regressor. We normalize the mean importance with *softmax* function, the final prediction of a datapoint $\boldsymbol{x}_i$ is given by:

$$\hat{y}_i = \sum_{v \in \Gamma_i} w_v \hat{y}_{vi} \quad \text{where} \quad w_v = \frac{\exp(\hat{y}_{vi} \bar{f}_v)}{\sum_{v \in \Gamma_i} \exp(\hat{y}_{vi} \bar{f}_v)} \tag{1}$$

where $\bar{f}_v$ is the mean feature importance of source v, and Γ_i is the set of sources that are present in $\boldsymbol{x}_i$

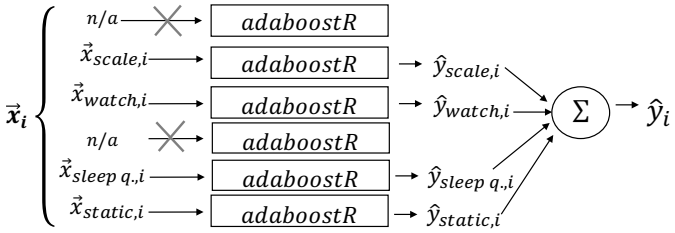


Fig. 5. Schematic of the ensemble model. A testing datapoint may not contain features from every source, depending on the availability of features, source-specific regressors make individual predictions, and are weighted by w_v to give a final prediciton. In this example, data from aura and the daily fatigue question are not present.

Ensemble of AdaBoost Regressors (AdaboostR). We use an ensemble of source-specific AdaBoost Regressor (AdaboostR) [9], where each regressor is trained independently on data of a specific source, namely the aura, scale, watch, or static (i.e. features from the screener survey). The component regressors have the same initializations but use different sets of features, resulting from independent feature selections. At testing phase, each regressor makes prediction using its associated subset of features from the datapoints, sources which are not present are ignored and no prediction is made. Finally a weighting function is used to average all the predictions. As illustrated in Figure 5, a major advantage of using an ensemble is that datapoints with different missing modalities can be considered.

The component regressor AdaboostR itself is an ensemble method. It fits predictors sequentially on a given dataset by setting weights to the predictors and the datapoints such that the subsequent predictors focus more on difficult cases. The choice of AdaboostR as the component regressor is driven by its accurate and robust performance, as will be discussed in later sensitivity analysis. Throughout the paper, we use AdaboostR to refer to the ensemble of AdaboostR.

Ensemble of Gaussian Mixture Regression Models (GMR). We compare the ensemble of AdaboostRs with an ensemble of Gaussian Mixture Regression Models (GMR) with diagonal covariance matrices. The latter choice is tied to our later considerations of model adaptations using a Maximum A Posterior (MAP) method with GMR.

Gaussian Mixture Regression is a technique developed for multivariate nonlinear regression modelling, it constructs a sequence of M Gaussian mixtures for a joint density of the data, and then derive conditional density and regression functions from each mixture [31]. Although in other texts 'mixture' and 'component' is used interchangbly, to avoid confusion, we refer to Gaussian mixture components g as 'mixtures', and the Gaussian mixture model itself as a 'component' of the ensemble model. We use GMR with $M = 3$ in our investigation.

The ensemble of GMRs is constructed in the same way as for AdaBoostR, where each GMR is trained independently per source using the Expectation Maximization (EM) algorithm. To initialize the EM algorithm, we use the k-Means++ algorithm to set initial means and variances of each GMR mixture. One important distinction from the ensemble of AdaBoostRs is that, predictions from component GMRs are combined using weighting function (Eq. 1) but instead of using the mean feature importance per source, we use the confidence statistics generated by the GMR models to weight the different component GMR per sample. In effect this means the weightings are dynamic and are computed on a case by case basis.

5.2 Model Adaptation

We explored ways of model adaptation of the above-mentioned universal prediction models (trained on data from all users), so that they may be trained on user-specific data give better generalization performance.

MAP-Adapted GMR. We consider a model adaptation method developed in [27] for speaker verification using Adapted Gaussian Mixture Models (GMM). Although originally intended for classification of speakers, this method satisfies our need for adapting to data specific to an individual in order to achieve better generalization. As we are considering regression, we adopt the formulation to Gaussian Mixture Regression (GMR) models [31] instead.

The main idea of MAP-Adapted GMR is to adopt the parameters of the GMR using Maximum A Posterior (MAP), by modifying the Maximization step in the EM-algorithm according to new observations, which in our case refer to the user-specific data. Each universal GMR would be modified that the posterior probabilities of the unimodal gaussian component given the new observations are maximized. As this is a non-iterative process so once the new sufficient statistics are calculated, the adaptation only needs to be performed once for each new set of observations.

The adaptation procedure is as follows: We first use data from the training set to produce an ensemble of universal GMR models, each with weight, mean and covaraince matrix parameters $\lambda_v(w, \mu, \Sigma)$. Then we consider user u from the testing set, and prepare his/her first T observations to be used for adaptation, $X_u = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_T\}$. For gaussian component g of each GMR λ_v, the component's posterior probability given the observation $\boldsymbol{x}_t$ is (we omit the v indexes for clarity in the following equations):

$$p(g \mid \boldsymbol{x_t}) = \frac{w_g p_g(\boldsymbol{x_t})}{\sum_{k=1}^{M} w_k p_k(\boldsymbol{x_t})} \tag{2}$$

Using the posterior probability of the component g, we then can calculate sufficient statistics for updating the weight, mean and variance:

$$n_g = \sum_{t=1}^{T} p(g \mid \boldsymbol{x_t}) \;, \quad E_g(\boldsymbol{x}) = \frac{1}{n_g} \sum_{t=1}^{T} p(g \mid \boldsymbol{x_t}) \boldsymbol{x_t} \;, \quad E_g(\boldsymbol{x x'}) = \frac{1}{n_g} \sum_{t=1}^{T} p(g \mid \boldsymbol{x_t}) \boldsymbol{x_t x_t'} \tag{3}$$

The updated parameters of the MAP adapted GMM are calculated using α as follows:

$$\alpha_g = n_g / (n_g + \tau) \tag{4}$$

$$\hat{w_g} = [\alpha_g n_g / T + (1 - \alpha_g) w_g] \gamma \quad , \quad \hat{\boldsymbol{\mu_g}} = \alpha_g E_g(\boldsymbol{x}) + (1 - \alpha_g) \boldsymbol{\mu_g} \;, \tag{5}$$

$$\hat{\boldsymbol{\sigma_g}}^2 = \alpha_g E_g(\boldsymbol{x x'}) + (1 - \alpha_g)(\boldsymbol{\sigma_g^2} + \boldsymbol{\mu_g^2}) - \hat{\boldsymbol{\mu_g}}^2 \tag{6}$$

where γ is a normalization coefficient such that $\sum_g^M \hat{w}_g = 1$ and τ is a relevance factor of the original model, a higher τ means a greater weighting will be given to the universal model.

Translation Using Residual Error. The second approach consists of a very simple transformation and was motivated by insights when analyzing prediction results given by the universal models by adjusting *residual errors*. This is motivated by the observation that for most predictions given by the universal models, the predictions were consistently off by some similar value for datapoints belonging to the same user. The simple transformation to adjust this is, to use the bias for the user's first completed weekly questionnaire, and to apply this bias onto all subsequent predictions given to that user. In effect, we apply a scalar translation of the prediction according to the residual error of the first week. Subsequent weeks' predictions are given by:

$$\hat{y}_i' = \hat{y}_i + b \tag{7}$$

where b is the residual error for the first week reported by the user which y_i is associated with.

5.3 Model Selection and Evaluation

Here we describe how we select features, fine-tune for hyperparameters, and evaluate the model performance.

Feature Selection. We have a total of 248 features and 1508 examples. We split users in the dataset randomly for training and testing, so that roughly 80% of the datapoints are in the training set and 20% in the testing set. We further divide the training dataset into source-specific subsets. Feature selection is done using the importance vector of AdaboostR per source. Each subset of data is fed into an AdaboostR, which generates an importance vector through averaging the importance vectors provided by its base regressors. We select features with a feature importance above the mean importance iteratively until the number of selected features are under a certain threshold D, per source. D is determined through grid search for each model in each prediction task.

Hyperparameter tuning. The hyperparameters (feature threshold, window length, window end-day, number of windows, imputation thresholds) are selected through a grid search process where integer values within a reasonable range for each hyperparameter were tested. We use leave-one-user-out cross validation such that such that the average mean absolute error is minimized.

Evaluation. As baseline, we also report results on a hypothesized model which always predicts the global mean score for each prediction task. We report performance achieved by the considered models using the following metrics:

- *Mean Absolute Error (MAE).* To assess the absolute difference between predicted values and ground truth across the entire population. For FSS this is with reference to a score that ranges from 1 to 7, and for EQ-5D-5L a score that ranges from 0 to 1.
- *Pearson correlation r*. To assess the correlations r between predicted values and ground truth across the entire population, where p-value means the significance.

6 PREDICTNG FATIGUE AND HEALTH STATE USING UNIVERSAL MODELS

In this section, we present prediction results of the universal ensemble models of AdaboostR and GMR for the prediction tasks of FSS-M and EQ-M, and FSS-W and EQ-W.

6.1 Predicting FSS-M and EQ-M

Table 2 and 3 shows the performance of the prediction models in predicting mean fatigue and health state respectivel, both ensemble models of AdaboostR and GMR outperform baseline and achieve signifcant and strong correlation with ground truth.

Figure 6 shows a comparison of the MAE distribution across tested users for both models. In both tasks, the generalization performance of the ensemble models closely match each other, with the 75th-percentile for the FSS-M task averaging at 1.41 points in a 1-to-7 scale, and that of the EQ-M task averaging at 0.089 on a 0-to-1 scale.

Through grid search, we find that the threshold for number of features D should be in the 18 to 20 range for both tasks. The consequence of this large threshold per source is mainly explained by its effect on limiting the number of static features sourced from the screener questionnaire. For sources such as the daily sleep quality question, the effect of having a large D has no effect since it only has 6 features. However, this becomes a significant criteria for the static source, which has 82 features. The observation that the models would perform better when considering more static variables makes sense as predicting the mean scores relates to capturing more stable relations that may be better captured by the static variables.

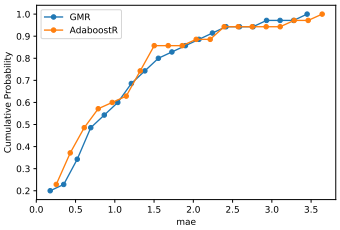
	Baseline	AdaboostR	GMR
mae	1.05	0.95	1.00
r	-	0.57*	0.34*

Table 2. FSS-M, * p<0.05

	Baseline	AdaboostR	GMR
mae	0.11	0.086	0.084
r	-	0.63*	0.72*

Table 3. EQ-M, * p<0.05

(a) FSS-M



(b) EQ-M

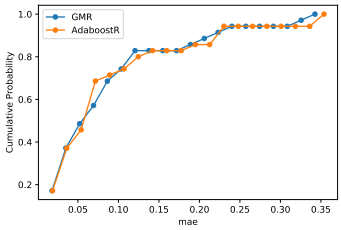


Fig. 6. CDFs of MAEs obtained in predicting FSS-M and EQ-M.

6.2 Predicting FSS-W and EQ-W

Table 4 and 5 shows the performance of the prediction models in predicting weekly fatigue and health state respectively. AdaboostR consistently achieves better MAE than the baseline and GMR in both tasks, outperforming GMR by 17.4% on average.

Figure 6 shows a comparison of the per-person MAE distribution across tested users for both models. AdaboostR is seen to achieve better generalization performance, this is especially true for the FSS-W task, where GMR is seen to result in a large error of 3.5 points in the worst case, while AdaboosR results in MAE of 2 points in the worst case.

For these tasks, we observe setting the number of features D as lying between 9 to 11 work best. The above results are computed with the number of time series windows consider set as 2, time series length as 10. For FSS-W prediction, the best model for AdaboostR only consists of the subset of source regressors from watch, daily fatigue question, and static.

	Baseline	AdaboostR	GMR
MAE	1.26	0.99	1.20
r	-	0.65*	0.25*

Table 4. FSS-W, * p<0.05

	Baseline	AdaboostR	GMR
MAE	0.11	0.091	0.11
r	-	0.61*	0.56*

Table 5. EQ-W, * p <0.05

7 MODEL ADAPTATIONS

For model adaptations, we compare two approaches: MAP-Adapted GMR with T user-supplied labels (*Adapted-GMR (T)*), and residual-error shifted AdaboostR (*Adapted-AdaboostR)*. The model performances are compared in Table 6 and 7.

(a) FSS-W (b) EQ-W

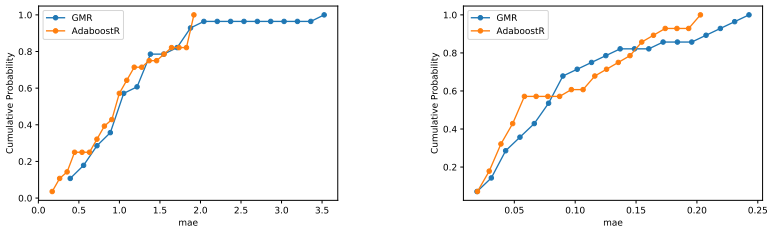


Fig. 7. CDFs of MAEs obtained in predicting FSS-W and EQ-W.

	Baseline	Universal-AdaboostR	Universal-GMR	Adapted-AdaboostR	Adapted - GMR(1)
MAE	1.26	0.99	1.20	0.51	0.97
r	-	0.65*	0.25*	0.91*	0.65*

Table 6. FSS-W predictions using universal and adapted models using 1 user-supplied datapoint,* p<0.05

	Baseline	Universal-AdaboostR	Universal-GMR	Adapted-AdaboostR	Adapted - GMR(1)
MAE	0.11	0.091	0.11	0.052	0.085
r	-	0.61*	0.56*	0.87*	0.80*

Table 7. EQ-W predictions using universal and adapted models using 1 user-supplied datapoint, * p<0.05

7.1 Adapted Gaussian Mixture Regression

Despite the small number of datapoints being available per person, we observe that the MAP-Adapated GMRs works well in giving better performance than the universal GMRs. The degree to which performance is improved is observed to be closely related to the number of adapted datapoints per individual T. We varied the number of datapoints from 1 to 10 and found that the MAE decreases monotomically as the number of datapoints T increases. As shown in Figure 8, on the whole the MAE distributions shifts more to the left as more datapoints are being adapted, for both FSS-W and EQ-W. In both cases, only using 1 datapoint per person is enough to improve the MAE distribution significantly, for FSS-W, this means a drop of 19% of MAE from 1.20 to 0.97, and for EQ-W, a drop of 23% from 0.11 to 0.085.

In addition, we also consider the changes in MAE per person using the universal versus MAP-adapted GMR. At $T = 2$, 4 out of 28 users experienced a small increase in MAE when using the MAP-adapted GMR, averaged at 3.5%. In the worst case, one user's MAE increased by 5%. Upon closer inspection we find that this user's first 2 weeks of data only consisted of static and Aura data, so only MAP-Adapted GMRs are available for these sources. This user started using all other modalities in the subsequent weeks, but predictions were made for those sources using GMR components which are not adapted to his/her data.

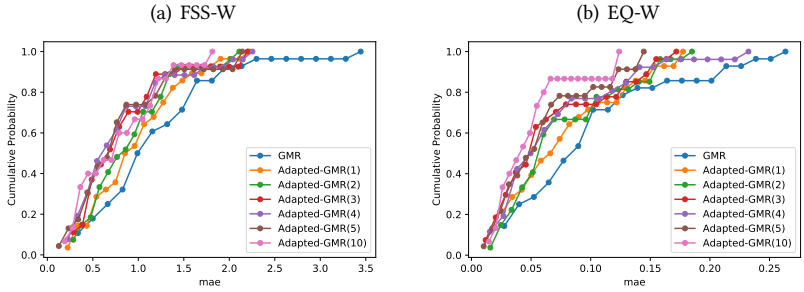


Fig. 8. CDFs of MAEs when varying adaptation threshold in GMR models for predicting FSS-W and EQ-W.

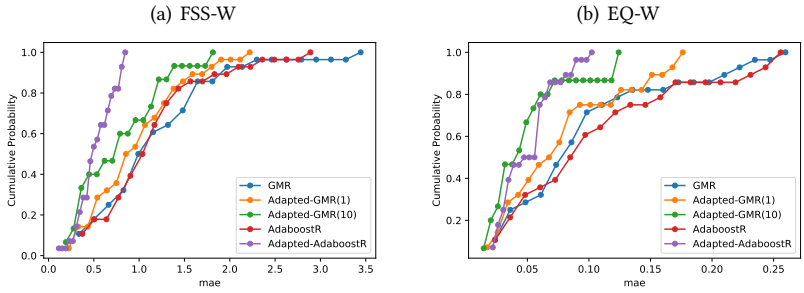


Fig. 9. CDFs of MAEs when varying adaptation threshold in GMR models for predicting FSS-W and EQ-W.

7.2 Translation of Residual Error

Here we apply a simple scalar translation of the prediction according to the residual error obtained in the the first week, applied onto predictions made by the AdaboostR ensemble model. We note that this simple operation yields a powerful reduction of MAE in both FSS-W and EQ-W case. For FSS-W, MAE is reduced by 49%, and for EQ-W, MAE is reduced by of 53%.

In Figure 9 we compare the universal models as well as the two adaptation methods, simple translation and adapted GMR. We also included two models of adapted GMR with 1 and 10 datapoints adapted per person to represent a scarce and abundant data situation. For FSS-W, it is clear that the simlpe translation is enough to outperform all other considered model. For EQ-W, only the Adapted-GMR which takes 10 datapoints per individual (with MAE 0.049) could match the performance given by the simple translation approach (MAE 0.052). Although this presents possibilities that the error could be reduced even further if we keep on increasing the number of adapted datapoints, this situation is unlikely in the future, given that in the current study the mean number of weekly questionnaires completed is under 10. However, this result also highlights that using the simple transformation is a pragmatic approach which works effectively in personalizing the model to individual users.

To ensure this translation is effective, we also consider the changes in MAE per person before and after applying the translation. Out of 28 subjects tested, 3 subjects experienced a minute increase in MAE that is up to 0.55%, and 2 subjects experienced small increase in MAE up to 2.3%. Upon closer inspection, we believe that reasons for the worsening performance for these individuals are due to their longer time series length. The 2 subjects have a

Table 8. Model selection.

Item		Alternatives explored
Data with Imputed Gaps		Filling gaps of 1, 4, 7 days
Multimodal Fusion		Single regressor , ensemble
Component Models		AdaboostR, GBRT, LASSO SVR (Linear), SVR (RBF)
Data Source		All sources, Device only, Device + Static, Daily Questions + Static,
Data Grouping		Device-level, Modality-level
Window Selection	Window Length	5 to 14 days
	Number of Windows	1 or 2 windows
	Window end point	-5 to 5 days before weekly survey fill-in

	Baseline	Ensemble	Single
mae	1.05	0.95	1.08
r	-	0.57*	0.29*

Table 9. Predicting FSS-M with single or ensemble regressors, * p<0.05

	Baseline	Ensemble	Single
mae	0.11	0.084	0.089
r	-	0.72*	0.58

Table 10. Predicting EQ-M with single or ensemble regressors, * p<0.05

mean time series length of 16.1weeks, 85% longer than the rest of the test set. It is possible that the adaptation could benefit from a recaliberation.

8 SENSITIVITY ANALYSIS

In arriving at the results presented in previous sections, we performed a number of investigations varying a number of model formulations and different hyperparameters, as documented in Table 8

We perform grid search to obtain optimal values using cross validation. In cases where the performance of of the alternatives are not significantly different, we opt for choices which are the simplest and easiest to interpret. For instance in the case of choosing the window end point, we choose this as 0, i.e. considered window ending on the date of filling in the weekly questionnaire. In the following we present some comments relating to the prediction tasks of FSS-M and EQ-M:

Multimodal fusion. As an alternative to an ensemble of regressors, we also considered employing a single regressor with concatenated features from the different modalities, this essentially represents an early multimodal fusion. One downside to this approach is that there cannot be any null values for any feature within a single datapoint, as was the case in ensemble, to overcome this it is required to select features which would be both important and abundant together. Table 9 and 10 show comparisons between the formulations. For predicting FSS-M, a single regressor predicts worse than the best ensemble and the baseline. For predicting EQ-M, although the inccrease in MAE is small, the correlation found is not as strong or significant (with p = 0.06).

Subset of data. We also consider subsets of source-specific regressors that is included in the ensemble model. We compared results of considering devices-only, devices + static, and all (devices + daily quesitons + static), and

we observed no great difference in the resulting models' MAEs, variations are within 5%. However, using devices data alone does significant harm the correlation coefficient, with an average decrease of 57% in r.

9 DISCUSSION

We present the discussion of the results, their implications and limitations of our study.

9.1 Discussion

The Data. *Use of Derived Data.* Like many related studies, our analysis builds on derived data from lower-level activity inference models that are fairly mature in the industry; as seen in [36, 37], the authors have relied on inferences such as walking, running, sleep stages build on prior classifiers. The derived data is inferred using commercial-quality models by Withings, and we have verified through discussions that they have validated their models in similar fashions to those paper in the literature. (i.e., controlled user trials) at a similar scale. In addition, these devices are used by millions of people on a daily basis - an added check since other methods have not been tested against. Further, we highlight that our dataset, being a mixture of raw information (e.g. weight, height) as well as derived data, represents an increasingly popular form of data useful for wellbeing research. However, we do acknowledge that the accuracy of the derived data might not be perfect, since these devices are currently intended to be used as 'general wellness' as opposed to regulated medical device [5]. We are also limited by the fact that the derived data are only available at daily levels and cannot investigate more fine-grained behavioral patterns, e.g. steps during different times of the day. Nonetheless we would like to investigate the possibility of using such derived measurements as proxies for personal disease monitoring at home by the patient. In a preprocessing stage we also set sensible criteria on the derived data to remove unreasonable measurements, e.g. measured bed-in time should be between 6pm and 6am, which inevitably creates a possibility for the rejection of genuine measurements.

Missingness. The resulting dataset contains many gaps due to the absence of data. Although the patients were conscious participants of the study, data was collected effectively in the wild in the sense that there was no requirement about device usage or daily/weekly questionnaire participation. For device measurements, participants' usage is a key factor leading to missingness - participants may not use the devices every day, they may not use all 3 devices on the same day, or their use of a single device may not enable all measurements to be derived for the day (e.g. a patient may wear smart watch during the day but not during sleep). A small number of these gaps is due to the removal of unreasonable derived values in a bid to ensure data quality. Low participation is seen particularly in the Daily questions, which was not attempted at all by half the participants. The missingness gives rise to issues such as a limited number of training examples and missing modality problem, but we have also found it to be useful to take into account of missingness as an indicator for patient's usage patterns. This has shown to be a valuable feature that is retained by the feature selection algorithm we employed with AdaboostR, and we also found that if we consider data with imputed gaps without the usage features, the model performance is not as good as when we take into account usage features. This means that some cases, the actual measurement does not matter as much as whether a measurement was actually made.

The daily questions. We note that these daily questions are not validated clinical instruments and are thus not necessarily reliable and robust measures for tracking fatigue and sleep quality. Contrary to expectations, data collected through these modalities do not always improve prediction performance, in fact for FSS-W the best model considers a subset of regressors that ignores these daily questions. One reason for this is that the data collected here is noisy and incomplete. It is not uncommon that we observe patients reporting No or Low fatigue in the daily questions just a few days prior to reporting in the FSS that they experience high fatigue with a score above 6. Since the patients fill in these daily questions as much as they please, we are left with incomplete data

which makes us uncertain about the usefulness of these daily questions, however this also means that sometimes the action of whether answering the daily question or not (captured by the usage feature) is more informative than the actual answer.

Ensemble Model. Under the ensemble model formulation, a datapoint only needs to have at least data from one time-varying source in order to be considered valid. This greatly increases the flexibility in the model, as the tracking of fatigue and health status can be thus extended to users who own and use different subsets of the wellness devices, a realistic situation if our modelling framework is applied in the commercial market in the future.

Adaptations. The simple adapted ensemble model of Adaboost Regressors is able to significantly reduce errors by 51% on average and improves generalization performance. We believe that the simple adaptation is a pragmatic approach given the sparsity of data. By empirical observation it is clear that adjusting the first residual error does not describe all of the individual prediction errors, but it is an effective approach and under no circumstances do we see a great (beyond 2%) increase in error in any individual. Predictions from this adapted model show good correlations with the ground truth on a population level and on an individual level for both FSS-W and EQ-W.

Alternative formulations. In terms of performance, ensemble models are also seen to perform better than single classifiers which take in imputed data, showing that our framework is a valid choice for handling multimodal data. Although we presented the best results from the best ensemble model resulting from hyperparameter tuning in previous sections, our sensitivity analysis also shows that the performance resulting from different hyperparameter settings are not far off. This supports that our prediction results are not the products of extreme hand-engineered solutions but could be possibly matched by similar models which might be varied if deployed in the future.

9.2 Implications of Results

Altogether, our results confirm the feasibility that weekly reported scores of Fatigue and Health State can be accurately and regularly tracked in MS patients. One implication is that, now patients may be able to employ such a model to track themselves for long periods of time at weekly intervals, simply by using a general model that calibrates once. Given that the disease does progress we might expect that the patients could re-calibrate every 4 months or so, in order to improve personal adaptations. Although patients do have to be equipped with smart devices, but once this is set up and calibrated, the burden of measuring week-to-week fatigue scores can be transferred to the devices and the patients would be able to use this information in conjunction with other summaries provided by the devices to develop a holistic view of their conditions. Being able to obtain at ease such descriptions of fatigue in terms of a clinically-validated equipment instead of arbitrary personal metrics also meant that the patients can share with their doctors more useful and more objective measures of their fatigue.

9.3 Limitations

While the current study provides evidence that ubiquitous sensing with connected wellness devices may help track and predict fatigue levels and health states of MS patients, there are a number of limitations.

The Data. Although the number of patients investigated are comparable to most MS studies, the size of our dataset is small and the demographics of patients investigated are limited (mostly female RRMS patients). Our collected data does not provide measurements more fine-grained than daily levels, it is possible that morning/afternoon fluctuations of activities could be important. Gait is the only proxy for activity or exercise taken by the patients here, which may also be too narrow. In addition, we also have not collected any data that is immediately related to the emotional wellbeing of the MS patients, despite that this could be a major factor in a patient's perceived levels of fatigue. Although the screener surveys were done to collect background variables from the patients, it is

probable that these variables (e.g. drug choice) could change over the course of the study, but this was not being tracked. For most background variables, only binary indications are given, this does not give the full picture and in future studies we believe a smaller number of background variables may be collected but at greater detail (e.g. years since diagnosed with other diseases instead of whether other diseases are present).

Ensemble Modelling. Although our final results demonstrate low levels of MAEs for both FSS and EQ-5D-5L scores, in our ensemble model we have only considered homogeneous settings for each component regressor (except for feeding in source-speicific data). It is possible that more complicated forms of the ensemble, where each component is fine-tuned independently, including the use of different types of regression models per source, could lead to better results which capture inter-modality variations.

To pave way for future work in this direction, future studies should seek to address the limitations in data collection. It would also be interesting to have some incorporation of measurements of patients' emotional state, for instance, through use of smartphone data or Ecological Momentary Assessment (EMA) methods. It would also be useful to collect information relating to environmental effects (e.g. local weather, humidity information), which had been shown in MS literature to be related to symptom triggers. In order to better adapt to personal changes, low-user-burden ways of calibrating the models could be investigated, e.g. asking for a binary indication of whether this week feels more fatigued than last week.

10 RELATED WORK

In recent years, increasing interest has been placed on using ubiquitous sensing technology to aid Multiple Sclerosis disease management. Digital and remote control technologies, enabled by the availability of cheap and ubiquitous sensors, can help over come previous limitations in data collection at low cost. A lot number of pilot studies have been carried out to explore the feasibility of deploying ubquitous sensing in MS patient's natural environments so that longitudinal assessment of their symptoms could be performed. The focus of many such studies tend to be in exploring the feasibility of data collection using ubiqutious sensing and analyzing the correlations. [22] explored using real-time depth sensors to identify gait problems and falls in 21 MS patients, and found that depth sensors in home could gather real-time gait parameters but not for detection of falls. [30] used activity sensors to gather information about the level of physical activity carried out by 11 MS patients, and correlated the tracked daily physical activity fluctuations to disability changes in MS patients. [21] performed a study of a larger scale, with 248 MS patients being handed FitBit activity trackers to collect actvitiy data, the authors analysed correlations found in the tracked data and reported data. Certain technologies have already been developed for MS monitoring and management which are alternative or complementary to traditional in-clinic approaches. *MS Mosaic* is a smartphoen app that allows MS patients to track their symptoms over time by manually reporting their symptoms (e.g. Faituge)and also integrating health and fitness data available on smartphones to monitor symptom triggers. *Floodlight* is also another smartphone apps which tracks changes in MS over time through 'active tests', where patients are asked to perform daily simepl tasks on their smartphone, as well as passive monitoring through smartphone health data. [3] have reported the results from the same dataset studied in our work, but with basic analytic results about correlations between expected fatigue changes and behavioral measurements. In general, most of these related studies have mostly relied on ground truth symptom measurement which had not been previously clinically validated, as far as we know, no technology so far have been developed for tracking fatigue and health state as measured in terms of the FSS and EQ-5D-5L.

Closely related to our work are also studies which have utilised ubiqutious sensing in monitoring symptoms for other diseases, many of which have been carried out to reproduce predictions for clinically-validated self-report instruments in the concerned disease. For example, [36] develops a prediction system that tracks schizophrenia symptoms based on a standard instrument using passive sensing from mobile phones, they were able to accurately predict reported schizophrenia scores using Gradient Boosted Regression Trees (GBRT). [37] proposed a new

approach in predicting depression using passive sensing data from college students' smartphone and wearables through the use of a proposed set of symptom features, they used generalized linear mixed model (GLMM) to predict self-reported depression scores and found correlations between their proposed symptom features and depression scores. Other than disease monitoring, a number of studies have also been carried looking into monitoring of more general wellness indicator, since this could be applied to the general non-clinical population, the sample dataset sizes studied could be much bigger, therefore also allowing more advanced techniques to be applied (e.g. deep neural networks). [35] analyzes multimodal time-series data and predicts the ability to achieve weight objective for users of smart connected devices using deep long-short-term memory architectures.

11 CONCLUSION

As a lifelong debilitation disease that affects millions of people worldwide, MS needs to be better understood by researchers and better monitored by patients. Having large-scale longitudinal data of MS patient's condition furthers this goal, and in this study we carried out the first investigation into methods to obtain such data through ubiquitous sensing at ease. In particular we focused our tasks on data relating to MS patient's fatigue and quality of life, through use of two widely used instruments by the MS community, Fatigue Severity Scale (FSS) and EQ-5D index. We conducted a study to collect behavioral, physiological device data and self-reported data from 198 MS patients, using connected wellness devices over 6 months. In our investigations, we proposed an ensemble regression models which can cope with the data's missingness, as well as adaptation techniques to further improve generalization performance. Our models are able to achieve good prediction performance for the tasks considered, namely predicting a per-participant mean reported score and a per-week per-participant score for each metric. We find that with the universal mode performance are in line with acceptable instrument errors, for FSS (SEM 0.7) we report MAE 0.99 and for EQ-5D (SEM 0.093) we report MAE 0.091. Adapted models improve these results further (FSS: MAE 0.51, EQ-5D: 0.052). These promising results in our dataset show the feasibility in continuous and unobtrusive tracking of fatigue and health state, and potential for future replications into larger-scale replication studies, which has positive implications for supporting MS patient disease management and clinical research.

REFERENCES

[1] 2018. Withings. https://www.withings.com. Accessed: 2018-11-08.
[2] B. E. Aouizerat, C. A. Miaskowski, C. Gay, C. J. Portillo, T. Coggins, H. Davis, C. R. Pullinger, and K. A. Lee. 2010. Risk factors and symptoms associated with pain in HIV-infected adults. *J Assoc Nurses AIDS Care* 21, 2 (2010), 125–133.
[3] Sourav Bhattacharya, Alberto Gil C. P. Ramos, Fahim Kawsar, Nicholas D. Lane, Lynn M. Gionta, Joanne Manidis, Greg Silvesti, and Mathieu Vegreville. 2018. Monitoring Daily Activities of Multiple Sclerosis Patients with Connected Health Devices. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp '18)*. ACM, New York, NY, USA, 666–669. https://doi.org/10.1145/3267305.3267682
[4] A. Bisecco, G. Caiazzo, A. d'Ambrosio, R. Sacco, S. Bonavita, R. Docimo, M. Cirillo, E. Pagani, M. Filippi, F. Esposito, G. Tedeschi, and A. Gallo. 2016. Fatigue in multiple sclerosis: The contribution of occult white matter damage. *Mult. Scler.* 22, 13 (11 2016), 1676–1684.
[5] Tim Bradshaw. 2018. Tech and healthcare often struggle to sync. *Financial Times* (29 Jan 2018). https://www.ft.com/content/9de679a4-04dd-11e8-9650-9c0ad2d7c5b5
[6] R. F. Brown, E. M. Valpiani, C. C. Tennant, S. M. Dunn, M. Sharrock, S. Hodgkinson, and J. D. Pollard. 2009. Longitudinal assessment of anxiety, depression, and fatigue in people with multiple sclerosis. *Psychol Psychother* 82, Pt 1 (Mar 2009), 41–56.
[7] ROBERT FERRARI and ANTHONY SCIENCE RUSSELL. 2010. Effect of a Symptom Diary on Symptom Frequency and Intensity in Healthy Subjects. *The Journal of Rheumatology* 37, 11 (2010), 2387–2389. https://doi.org/10.3899/jrheum.100513 arXiv:http://www.jrheum.org/content/37/11/2387.full.pdf
[8] E. Fogarty, C. Walsh, R. Adams, C. McGuigan, M. Barry, and N. Tubridy. 2013. Relating health-related Quality of Life to disability progression in multiple sclerosis, using the 5-level EQ-5D. *Mult. Scler.* 19, 9 (Aug 2013), 1190–1196.
[9] Yoav Freund and Robert E Schapire. 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. System Sci.* 55, 1 (1997), 119 – 139. https://doi.org/10.1006/jcss.1997.1504
[10] M. M. Goldenberg. 2012. Multiple sclerosis review. *P T* 37, 3 (Mar 2012), 175–184.

[11] M. Heine, I. van de Port, M.B. Rietberg, E.E.H. van Wegen, and G. Kwakkel. 2015. Exercise therapy for fatigue in multiple sclerosis. *Cochrane Database of Systematic Reviews* 9 (2015). https://doi.org/10.1002/14651858.CD009956.pub2

[12] Sverker Johansson, Anders Kottorp, Kathryn A. Lee, Caryl L. Gay, and Anners Lerdal. 2014. Can the Fatigue Severity Scale 7-item version be used across different patient populations as a generic fatigue measure - a comparative study using a Rasch model approach. *Health and Quality of Life Outcomes* 12, 1 (22 Feb 2014), 24. https://doi.org/10.1186/1477-7525-12-24

[13] Sverker Johansson, Charlotte Ytterberg, Jan Hillert, Lotta Widén Holmqvist, and Lena von Koch. 2008. A longitudinal study of variations in and predictors of fatigue in multiple sclerosis. *Journal of neurology, neurosurgery, and psychiatry* 79 4 (2008), 454–7.

[14] S. Johansson, C. Ytterberg, J. Hillert, L. Widen Holmqvist, and L. von Koch. 2008. A longitudinal study of variations in and predictors of fatigue in multiple sclerosis. *J. Neurol. Neurosurg. Psychiatry* 79, 4 (Apr 2008), 454–457.

[15] E. Kim, J. Lovera, L. Schaben, J. Melara, D. Bourdette, and R. Whitham. 2010. Novel method for measurement of fatigue in multiple sclerosis: Real-Time Digital Fatigue Score. *J Rehabil Res Dev* 47, 5 (2010), 477–484.

[16] G. Kobelt, J. Berg, P. Lindgren, S. Fredrikson, and B. Jonsson. 2006. Costs and quality of life of patients with multiple sclerosis in Europe. *J. Neurol. Neurosurg. Psychiatry* 77, 8 (Aug 2006), 918–926.

[17] L. B. Krupp, P. K. Coyle, C. Doscher, A. Miller, A. H. Cross, L. Jandorf, J. Halper, B. Johnson, L. Morgante, and R. Grimson. 1995. Fatigue therapy in multiple sclerosis: results of a double-blind, randomized, parallel trial of amantadine, pemoline, and placebo. *Neurology* 45, 11 (Nov 1995), 1956–1961.

[18] L. B. Krupp, N. G. LaRocca, J. Muir-Nash, and A. D. Steinberg. 1989. The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. *Arch. Neurol.* 46, 10 (Oct 1989), 1121–1123.

[19] Amy E. Latimer-Cheung, Lara A. Pilutti, Audrey L. Hicks, Kathleen A. Martin Ginis, Alyssa M. Fenuta, K. Ann MacKibbon, and Robert W. Motl. 2013. Effects of Exercise Training on Fitness, Mobility, Fatigue, and Health-Related Quality of Life Among Adults With Multiple Sclerosis: A Systematic Review to Inform Guideline Development. *Archives of Physical Medicine and Rehabilitation* 94, 9 (2013), 1800 – 1828.e3. https://doi.org/10.1016/j.apmr.2013.04.020

[20] Y. C. Learmonth, D. Dlugonski, L. A. Pilutti, B. M. Sandroff, R. Klaren, and R. W. Motl. 2013. Psychometric properties of the Fatigue Severity Scale and the Modified Fatigue Impact Scale. *J. Neurol. Sci.* 331, 1-2 (Aug 2013), 102–107.

[21] James McIninch, Shoibal Datta, Pronabesh DasMahapatra, Emil Chiauzzi, Rishi Bhalerao, Alicia Spector, Sherrie Goldstein, Liz Morgan, and Jane Relton. 2015. Remote Tracking of Walking Activity in MS Patients in a Real-World Setting (P3.209). *Neurology* 84, 14 Supplement (2015). arXiv:http://n.neurology.org/content http://n.neurology.org/content/84/14_Supplement/P3.209

[22] P. Newland, J. M. Wagner, A. Salter, F. P. Thomas, M. Skubic, and M. Rantz. 2016. Exploring the feasibility and acceptability of sensor monitoring of gait and falls in the homes of persons with multiple sclerosis. *Gait Posture* 49 (09 2016), 277–282.

[23] Nokia. [n. d.]. *Nokia Health Mate app, Your Activity Tracker and LIfe Coach User Guide.* Nokia.

[24] Mari Palta, Han-Yang Chen, Robert M. Kaplan, David Feeny, Dasha Cherepanov, and Dennis G. Fryback. 2011. Standard Error of Measurement of 5 Health Utility Indexes across the Range of Health for Use in Estimating Reliability and Responsiveness. *Medical Decision Making* 31, 2 (2011), 260–269. https://doi.org/10.1177/0272989X10380925 arXiv:https://doi.org/10.1177/0272989X10380925 PMID: 20935280.

[25] T. Pereira, C. Correia, and J. Cardoso. 2015. Novel Methods for Pulse Wave Velocity Measurement. *J Med Biol Eng* 35, 5 (2015), 555–565.

[26] R. Rabin and F. de Charro. 2001. EQ-5D: a measure of health status from the EuroQol Group. *Ann. Med.* 33, 5 (Jul 2001), 337–343.

[27] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. 2000. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing* 10, 1 (2000), 19 – 41. https://doi.org/10.1006/dspr.1999.0361

[28] L. A. Rolak. 2003. Multiple sclerosis: it's not the disease you thought it was. *Clin Med Res* 1, 1 (Jan 2003), 57–60.

[29] K. M. Schreurs, D. T. de Ridder, and J. M. Bensing. 2002. Fatigue in multiple sclerosis: reciprocal relationships with physical disabilities and depression. *J Psychosom Res* 53, 3 (Sep 2002), 775–781.

[30] L. Shammas, T. Zentek, B. von Haaren, S. Schlesinger, S. Hey, and A. Rashid. 2014. Home-based system for physical activity monitoring in patients with multiple sclerosis (Pilot study). *Biomed Eng Online* 13 (Feb 2014), 10.

[31] Hsi G. Sung. 2004. *Gaussian mixture regression and classification.* Ph.D. Dissertation. Rice University.

[32] N Téllez, Jordi Rio, Mar Tintorè, C Nos, Ingrid Galán, and X Montalban. 2006. Fatigue in Multiple Sclerosis Persists Over Time: a longitudinal study. *Journal of neurology* 253 (11 2006), 1466–70. https://doi.org/10.1007/s00415-006-0247-3

[33] B. van Hout, M. F. Janssen, Y. S. Feng, T. Kohlmann, J. Busschbach, D. Golicki, A. Lloyd, L. Scalone, P. Kind, and A. S. Pickard. 2012. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Health* 15, 5 (2012), 708–715.

[34] Mandy van Reenen and Bas Janssen. 2015. *EQ-5D-5L User Guide.* EQ-5D.

[35] Petar Veličković, Laurynas Karazija, Nicholas D. Lane, Sourav Bhattacharya, Edgar Liberis, Pietro Liò, Angela Chieh, Otmane Bellahsen, and Matthieu Vegreville. 2018. Cross-modal Recurrent Models for Weight Objective Prediction from Multimodal Time-series Data. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '18).* ACM, New York, NY, USA, 178–186. https://doi.org/10.1145/3240925.3240937

[36] Rui Wang, Weichen Wang, Min S. H. Aung, Dror Ben-Zeev, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A. Scherer, and Megan Walsh. 2017. Predicting Symptom Trajectories of Schizophrenia Using Mobile Sensing. *Proc.*

ACM Interact. Mob. Wearable Ubiquitous Technol. 1, 3, Article 110 (Sept. 2017), 24 pages. https://doi.org/10.1145/3130976

[37] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. 2018. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 43 (March 2018), 26 pages. https://doi.org/10.1145/3191775

[38] Weichen Wang, Gabriella M. Harari, Rui Wang, Sandrine R. Müller, Shayan Mirjafari, Kizito Masaba, and Andrew T. Campbell. 2018. Sensing Behavioral Change over Time: Using Within-Person Variability Features from Mobile Sensing to Predict Personality Traits. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 141 (Sept. 2018), 21 pages. https://doi.org/10.1145/3264951

[39] K. Wynia, B. Middel, J. P. van Dijk, J. H. De Keyser, and S. A. Reijneveld. 2008. The impact of disabilities on quality of life in people with multiple sclerosis. *Mult. Scler.* 14, 7 (Aug 2008), 972–980.

Appendix B

Conference Paper: Ubicomp 2018

This work was completed in collaboration with researchers from The Univesity of Edinburgh, University of Cambridge and Nokia Bell Labs, Cambridge. This was accepted for the December 2017 issue of IMWUT and was presented in the ACM Conference on Pervasive and Ubiquitous Computing (UbiComp) 2018. I also gave a conference presentation on this work at MobiUK 2018 held in Cambridge.

Multimodal Deep Learning for Activity and Context Recognition

VALENTIN RADU, The University of Edinburgh
CATHERINE TONG, University of Oxford
SOURAV BHATTACHARYA, Nokia Bell Labs
NICHOLAS D. LANE, University of Oxford and Nokia Bell Labs
CECILIA MASCOLO, University of Cambridge
MAHESH K. MARINA, The University of Edinburgh
FAHIM KAWSAR, Nokia Bell Labs and TU Delft

Wearables and mobile devices see the world through the lens of half a dozen low-power sensors, such as, barometers, accelerometers, microphones and proximity detectors. But differences between sensors ranging from sampling rates, discrete and continuous data or even the data type itself make principled approaches to integrating these streams challenging. How, for example, is barometric pressure best combined with an audio sample to infer if a user is in a car, plane or bike? Critically for applications, how successfully sensor devices are able to maximize the information contained across these multi-modal sensor streams often dictates the fidelity at which they can track user behaviors and context changes. This paper studies the benefits of adopting *deep learning* algorithms for interpreting user activity and context as captured by multi-sensor systems. Specifically, we focus on four variations of deep neural networks that are based either on fully-connected Deep Neural Networks (DNNs) or Convolutional Neural Networks (CNNs). Two of these architectures follow conventional deep models by performing feature representation learning from a concatenation of sensor types. This classic approach is contrasted with a promising deep model variant characterized by modality-specific partitions of the architecture to maximize intra-modality learning. Our exploration represents the first time these architectures have been evaluated for multimodal deep learning under wearable data – and for convolutional layers within this architecture, it represents a novel architecture entirely. Experiments show these generic multimodal neural network models compete well with a rich variety of conventional hand-designed shallow methods (including feature extraction and classifier construction) and task-specific modeling pipelines, across a wide-range of sensor types and inference tasks (four different datasets). Although the training and inference overhead of these multimodal deep approaches is in some cases appreciable, we also demonstrate the feasibility of on-device mobile and wearable execution is not a barrier to adoption. This study is carefully constructed to focus on multimodal aspects of wearable data modeling for deep learning by proving a wide range of empirical observations, which we expect to have considerable value in the community. We summarize our observations into a series of practitioner rules-of-thumb and lessons learned that can guide the usage of multimodal deep learning for activity and context detection.

Additional Key Words and Phrases: Mobile sensing, sensor fusion, multi-modal, deep neural networks, deep learning, context detection, activity recognition

Authors' addresses: Valentin Radu, The University of Edinburgh; Catherine Tong, University of Oxford; Sourav Bhattacharya, Nokia Bell Labs; Nicholas D. Lane, University of Oxford and Nokia Bell Labs; Cecilia Mascolo, University of Cambridge; Mahesh K. Marina, The University of Edinburgh; Fahim Kawsar, Nokia Bell Labs and TU Delft.

2474-9567/2017/11-ART157 $15.00
https://doi.org/10.1145/3161174

ACM Reference Format:
Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D. Lane, Cecilia Mascolo, Mahesh K. Marina, and Fahim Kawsar. 2017. Multimodal Deep Learning for Activity and Context Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 157 (November 2017), 27 pages. https://doi.org/10.1145/3161174

1 INTRODUCTION

The popularity of wearables, and mobile sensing devices in general, has given rise to a growing interest in complex sensing applications (e.g. user activity and context recognition), with such tasks already available on commercial wearables to track jogging [42], sleep [48] and even posture [45]. Common to these recognition tasks is their reliance on numerous low-energy small form-factor sensors (e.g., light detector, magnetometer, accelerometer, barometer, heart-rate). With each sensing modality carrying a unique perspective, combinations of multiple such sensing streams can boost detection quality and exceed their potential in isolation. Taking this approach, the Microsoft Band [48] determines when a user is asleep by combining heart rate levels with accelerometer data of wrist motion, while the MSP [13] distinguishes between walking and climbing stairs, which are relatively similar in acceleration patterns, with extra information from a barometer.

Majority of current multimodal sensing solutions rely on *shallow* classifiers, (such as Decision Tree, Random Forest, SVM) operating on independent features extracted from each sensing modality. These features are used to perform sensor fusion following two strategies: Feature Concatenation (such as in [9, 28]) that treat features uniformly irrespective of their sensing modality to produce a single feature vector for classification; and Ensemble Classifiers (applied in [22, 72]) in which outputs of classifiers operating only on features of one modality are blended together. However, an important challenge for activity and context classification is to integrate seemingly incompatible sensor types (consider fusing accelerometer data and camera frames). Because sensing modalities vastly differ by sampling rate, statistical properties and data types, standard approaches for model training struggle to merge the information available from these diverse sources. The key here is to not only extract discriminative features from individual sensors, but also to discover features that jointly use separate sensors streams to capture information neither has in isolation.

Deep learning [2, 14] presents a promising, much unexplored opportunity to combat this sensors fusion challenge. In an area of rapid innovation, deep learning algorithms have shown to be remarkably successful in unimodal applications, such as recognition of words [27], objects [35] and faces [66]. One of its defining characteristics is the ability to learn dense hierarchical networks that transform relatively raw forms of data into inferences (e.g., an activity class). These networks merge the roles of features extraction and classification stages present in shallow modeling methods (e.g. SVMs [5]) and replace the need for hand-engineered, task-specific features with layers of data representations that act as features, automatically learned directly from data. There is already building evidence suggesting deep methods could overcome current bottlenecks in learning cross-sensor features for routine detections. New training methods that leverage variation in information [63], multi-view representations [69], or modified autoencoders [52, 70] are able to fuse highly heterogeneous pairs of data types, such as text mixed with images [64] and audio linked with video [52, 60]. The resulting bi-modality deep models offer considerable accuracy gains in tasks like image captioning [63] and emotion recognition [15, 33, 43] (merging facial expressions with sound).

This paper presents a case study of adopting deep learning algorithms for multimodal human activity and context recognition, using sensor data collected with mobile devices. We investigate the sensor fusion approaches taken in both deep and shallow learning, and ask the following questions: *(i) how do the studied techniques fare with existing practices? (ii) under what circumstances (e.g. modeling task, data types) is it beneficial to apply deep learning? (iii) how the technique may be deployed?* These questions are examined under two approaches to multimodal learning.

The first, *Feature Concatenation* (FC), is attractive in terms of simplicity as a strategy though it is at risk of missing crucial intra-modality correlations. The second, a novel alternative that we term *Modality-Specific Architecture* (MA) is a deep learning specific technique that places emphasis on learning both intra-modality and cross-modality relations. Our empirical study spans four datasets representative of a wide range of activity and context recognition tasks in ubiquitous computing. For each dataset, we evaluate: (i) four deep learning techniques based on FC and MA, with Deep Neural Networks (DNN) and Convolutional Neural Networks (CNNs) as base classifiers; (ii) two shallow classifier techniques: Decision Tree, Random Forest; (iii) any available task-specific classifier. Our MA architectures, trained here on mobile sensing data, are adaptations of the architecture first proposed in [52] for speech detection integrating video and sound modalities.

Our results show that these deep network architectures exceed current machine learning solutions over a range of mobile sensing tasks (human activity recognition, gait recognition, sleep stage detection and indoor-outdoor detection). This consistently better performance demonstrates a *general-purpose* characteristic of deep neural network architectures across diverse multimodal detection tasks, even matching that of highly-engineered *purpose-built* methods designed for a specific detection task. In addition to comparing accuracies, we also investigate the computational overhead of these techniques, as well as document the various lessons learnt in evaluating our training framework. Our findings suggest wearable resource limits (such as energy) are not a barrier to the adoption of explored deep learning methods.

Key contributions of this research are:

- A systematic study of multimodal deep learning techniques applied to a broad range of activity and context recognition tasks. Our empirical results demonstrate the ability of feature representational learning to produce accurate results even across highly heterogeneous sensors under different settings.
- We study the Modality-Specific Architecture, a specific type of split-architecture deep learning never previously applied to wearable modeling tasks. Within this variety of deep learning, we are also the first to test this architecture in-conjunction with convolutional layers (relative to feed-forward fully-connected DNN layers).
- Summary of experiences into general rules of thumbs and lessons learnt for future adoption of multimodal deep learning techniques in mobile sensing research.
- A system resource feasibility study of the overhead imposed by multimodal deep architectures. Experiments with two mobile processors show that memory, battery and computational footprint of these detection algorithms are not a barrier to their adoption.
- Development of a framework [18] to support training and evaluation of these models on different multimodal sensing tasks.

2 SHALLOW LEARNING APPROACHES

Shallow methods are commonly used for multimodal activity and context recognition. The term *shallow* is used to contrast with alternate deep learning architecutres [2, 14]. We first summarize the challenges of multimodal fusion, followed by surveying the two commonly adopted multimodal modeling strategies:

(i) Feature Concatenation

(ii) Ensemble Classifiers

2.1 Challenges for Multimodal Sensor Fusion

Different sensing modalities carry information with various perspectives, which often complement each other, allowing for useful information gain. As such, leveraging a diverse set of sensors on mobile devices when available can only be beneficial to detection accuracy.

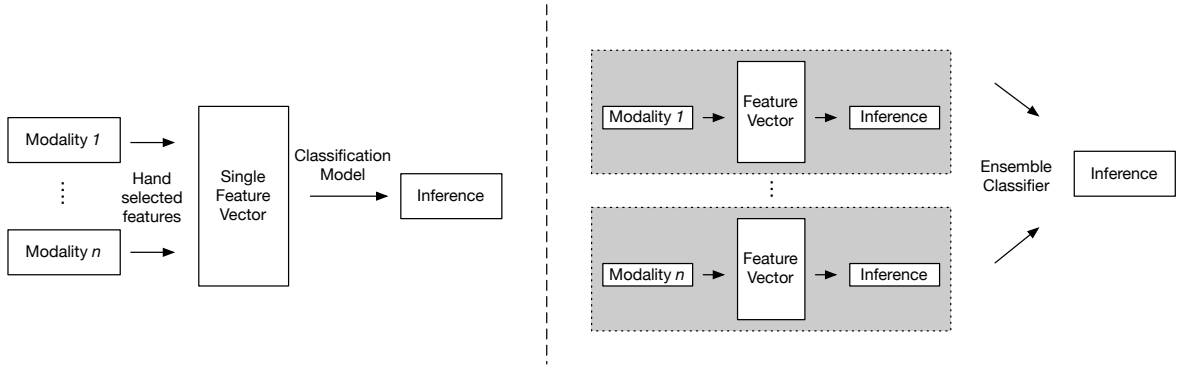
However, building appropriate models for multimodal fusion, i.e. models which fully leverage information contained within and across each sensor, is not trivial. The difficulty is mainly attributed to intrinsic differences between sensor data. Coming from different input channels with varying data types and sampling rates, each modality is characterized by distinctive statistical properties, representation and correlation structures. Typical models treat multimodal data as heterogeneous, resulting in ambiguous features and requiring special detection pipelines to deal with this multimodal data.

As a result of these differences, it is difficult to systematically recognize useful cross-modality relationships in addition to uni-modality ones. For instance, human speech can be perceived through three modalities: a stream of 2D-structured pixels (video), a time-series real-valued acoustic waveform (audio) and a sparse distribution of words (text). Cross-modality relationships which exist between low-level features across different modalities are highly non-linear and thus difficult to find. If we only consider video and audio, correlations are still highly non-linear and difficult to find from a mix of pixels and waveforms even if attempting by hand [52]. More generally, multi-modality data present even greater challenges since this difficult and time-consuming process needs to be performed for each sensor *pair*.

To illustrate this aspect in the context of multimodal sensing with mobile devices, we consider the following examples. A generic human activity recognition task utilizing the GPS, accelerometer and audio signals (e.g. indoor-outdoor detection) presents a range of challenges for modality fusion. Obviously, sampling rates differ substantially between sensor pairs: GPS signals change on time-scales of a minute, while accelerometer signal streams at sub-second rates (30Hz). This means that a GPS sample must be correlated to thousands of accelerometer readings; learning from such an imbalance distribution can easily degenerate typical models. Similarly, considering the GPS and audio signals, these are again very different. Even high sampling rate pairs like accelerometer and audio still need heavy filtering and preprocessing to align. Another example where these observations hold the same is the case of context understanding for indoor localization with smartphones. Typical sensing modalities for this task include inertial sensors (accelerometer and gyroscope) and WiFi scans, which come at different frequencies and in completely different formats (uniform data streams vs. uneven blobs of wireless environment observations). Integrating these sensing modalities is often done through heavy-engineered solutions, such as particle filters for indoor localization [58], but in most cases these are task-specific and constrained to a predetermined environment.

Several challenges exist in preprocessing sensor signals to comply with the input structure expected by a classifier. Generally, shallow classifiers operate on small size inputs which take advantage of the already distilled information presented to them in the form of features, specially hand-engineered for specific tasks – which can be an art on its own. Selecting the appropriate features may not be always intuitive to system designers even for a single modality; this difficulty is even more stringent with multimodal data as features have to complement each other across sensing modalities.

Nevertheless, shallow methods consider features as independent knobs, incapable of learning the key interdependent relations between sensing modalities, e.g. correlation in lip pose and motions between audio and visual data for speech recognition [52].



(a) Feature Concatenation (b) Ensemble Classifier

Fig. 1. Schematic of common approaches to shallow multimodal learning with (a) Feature Concatenation and (b) Ensemble Classifier models. For Feature Concatenation hand-selected features are extracted from each sensing modalities and concatenated into a single features vector as input to a classifier, while the Ensemble Classifier approach performs detections on each sensing modality independently to combine their estimations as final inference.

2.2 Strategies in Shallow Multimodal Learning

Existing sensor fusion strategies in multimodal activity recognition models can be categorized into two families, depending on whether sensor fusion occurs at a feature or at a classifier level. We call these (i) Feature Concatenation (FC), and (ii) Ensemble Classifiers (EC) respectively, schematically represented in Figure 1.

Feature Concatenation with Shallow Classifiers. With this strategy, hand-selected features from each modality are combined into a single feature vector presented to a classifier for detection across all features.

Various multimodal sensing systems have adopted this approach using different classifiers for diverse recognition tasks (SVM [9], AdaBoost [28], GMMs [44]). While some classifier types, such as ensemble learners like Random Forest [5], may do better than others at teasing out relationships between features the degree to which multimodal information is maximized is dependent on the quality of these hand-crafted uni- and cross-modal features. Often feature selection in concert with the extraction of a large number of candidate features for each sensing modality is attempted to automate this process (a technique adopted in systems like MSP [13]), though still bounded by the quality of selected features. In practice, Feature Concatenation can easily overlook inter-sensor relationships with the number of explored feature combinations limited by the curse of dimensionality [5].

Ensemble of Shallow Classifiers. On the other hand, the integration of modalities is done after separate classifiers operating on each sensor (modality) provide their estimation. These estimations provided by each sensing modality classifier are fused to yield an overall class estimation. A range of classifier fusion methods exists, including probability-based Bayesian fusion models and majority voting schemes (more details in [30]).

Like FC, variations of EC are also commonly adopted in multimodal activity models [30, 56, 72]. One key attraction is that available classifiers for each sensor type (tested and verified with other applications) can be readily adopted to undertake a new task on same sensing modality. In essence, this facilitates the creation of robust sensor-specific classifier generic to multiple tasks, while merging their results enables the evidence of each modality type to be considered before a final inference. However, a fundamental weakness of EC is that because fusion takes place so late a lot of potential information and cross-sensor relationships are already lost.

3 DEEP LEARNING TECHNIQUES

In this section, we discuss deep learning models used in multimodal activity and context recognition. We begin with an overview of existing methods, followed by a technical description of two kinds of multimodal deep learning models: *Feature Concatenation* (FC) and *Modality-Specific Architecture* (MA).

3.1 Overview

Deep learning models have been applied successfully to a growing number of detections with a single modality. Through their ability to learn feature representations directly from raw data (images, voice, text) rather than relying on domain-specific features and their hierarchical structure, deep models present a viable solution to overcome the challenges of multimodal sensing exposed above.

The structure of a generic deep learning architecture is presented in Figure 2. This consists of interconnected units that are grouped together in layers. Information propagates through layers, each performing transformations on their input as a function of internal feature representations, globally contributing to the final classification result. The first layer (input layer) accepts data in raw or lightly processed format, while the final layer (output layer) provides the class (e.g. categories of activity and context) according to the value of associated unit. Layers in between input and output layers are called hidden layers, because their values are not monitored, although essential to propagating information based on their layer parameters. Unit parameterization is automatically determined during training and depends on the layer type (e.g. whether it is convolutional or feed-forward layers) as well as training methods, pre-training and fine-tuning [14].

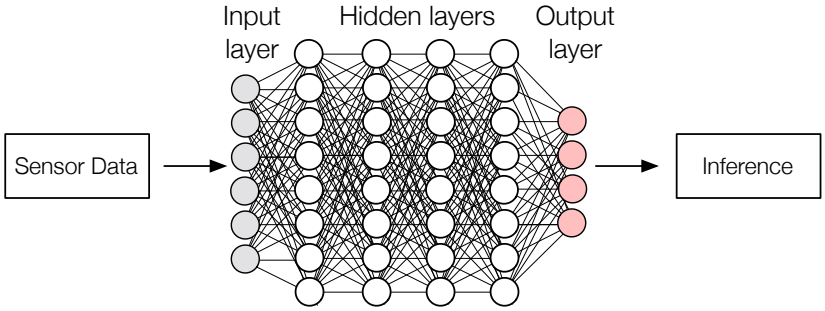


Fig. 2. Typical Deep Neural Network structure with feed-forward fully-connected layers.

Input Layer Representation. Deep architectures embody feature representational learning, with input layer taking values from raw sensor data, or lightly processed data.

Raw Data. Raw measurements, e.g. sensor readings, can be used as input directly.
Feature Selection. It is common for deep architectures to have a light preprocessing stage to change the dimensionality of the raw input signal (for example very sparse word vectors) in a process resembling features extraction in traditional classifiers, although this is a light and generic (valid across many tasks) data transformation process.

3.2 Multimodal Deep Learning

Deep learning architectures hold important properties which are advantageous to multimodal recognition tasks. As mentioned before, due to their feature representation learning, there is no need for a preprocessing phase to

extract domain-specific features from input data, this having two important consequences for multimodal data. First, custom layer representations can be trained to combine both uni-modal and cross-modal data from sensors. Second, the hierarchy in learned representations means that cross-modal relationships can be learned at both low- and high-levels of abstraction, corresponding to raw data and refined aggregated concepts respectively.

The same strategies for combining information from multimodal inputs as discussed for shallow classifiers can be applied to deep neural networks, based on the level where fusion is achieved: Feature Concatenation (FC) with concatenated inputs from multiple sensing modalities, and Modality-Specific Architecture (MA), with sensor-specific branches for each modality before fusion is achieved later in the network. We describe these two in more details below.

Essential to all machine learning classifiers is the training process. Various approaches to achieve training and tuning of multimodal architectures exist here, commonly built on the back-propagation algorithm. Back-propagation offers the flexibility of distributing gradients both on the joint section of the network, but also on the split section on each modalities branches. Flexibility of the network is not limited just to back-propagation, but also manifested in flexibility over data availability (or quality). As such, training algorithms specific to deep architectures facilitate robustness to missing modality inputs, which allows it to generate (or reconstruct) missing input modality from the available input by using the joint representation of relations between modalities. For example, auto-encoder algorithms, originally designed for uni-modal deep models to improve noise tolerance, are adapted to provide a type of tolerance to missing modalities [43, 52, 70]; this in turn, is understood to assist in discovering representations that are less prominent, but still discriminative, in power. One illustrative example of such an architecture is the image caption generation network, which can have its image representation layers improved (i.e., increased detection quality) through a training method that requires the model to generate both reasonable text and images when only an image is provided (with text related input layers set to a default state).

Many implementations of deep learning architectures with multiple modalities have been proposed in literature including Restricted Boltzmann Machine (RBM), CNN, DNN [51, 52]. We employ DNNs and CNNs under both sensor fusion strategies FC and MA, constructing 4 different architectures (FC-DNN, FC-CNN, MA-DNN, MA-CNN), which are described in the following sections.

3.3 Feature Concatenation Deep Learning

We refer to a commonly used approach in multimodal data integration using a deep classifier with concatenated modalities input as Feature Concatenation (FC).

In this approach, sensor fusion is performed right at the input layer by concatenating raw sensor streams (or lightly processed data) of multiple modalities, to achieve a single large input space. Data propagation pipeline inside the network proceeds as earlier described, performing a set of transformations on the concatenated input (Figure 2). This simple design allows easier training, since the model is less sensitive to hyper-parameter settings.

An important remark here is that feature representations inside the network have access to the whole space of sensing modalities (cross-modality information). However, previous work [52, 63] have shown that intra-sensor correlations (within the same modality) are stronger than inter-sensor ones (across multiple modalities). Since hidden layers in FC architectures are exposed to cross-modality information, it is harder to specialize them during training to extract the essential intra-sensor relations, so these get easily neglected. In addition, training an FC deep model is also problematic for an unbalanced mixture of inputs, as the units inside the network are easily dominated by those few proven modalities.

Feed-forward Deep Neural Network (FC-DNN). Feed-forward Deep Neural Networks are comprised of multiple stacked fully-connected layers, with the information passing from the input layer, starting as concatenated multimodal data and being transformed sequentially by each layer according to their internal feature representations and activations. Information flows in one direction through the network, which makes it easy to stack several hidden layers together. Class estimation is provided by the output layer as described before. In essence, modalities fusion is achieved at the input level by combining sensor streams into a joint input to propagate through the DNN.

Different activation function and regularization methods also bring their contribution to transformations propagated between hidden layers.

Convolutional Neural Network (FC-CNN). Similar to FC-DNN, sensing modalities are concatenated into a single input, which is interpreted with a Convolutional Neural Network in this case.

Convolutional Neural Networks have brought major leaps across many research areas [41]. Two layers are specific to this construction, Convolution layers and Pooling layers. Convolution layers are characterized by shared weights and biases as stacks of filters, much smaller in size than the input signal being convolved over. A stack of filters is trained to recognize different patterns (or features) no matter where these are encountered in the input space by sliding across the whole input. Although these filters are determined automatically in training, a good initialization strategy can determine non-overlapping behaviors of filters, each specializing in recognizing different patterns.

Convolution layers are typically followed by a Pooling layer, which reduces the size of the new representation constructed by filter activations. Pooling and Convolution layers can be stacked to produce more complex features, progressing in composition throughout the network. A common example comes from computer vision where first Convolution layers detect simple features like lines, points, colors, followed by other Convolution layers that activate for shapes, corners, edges and so on until by the end they recognize full body-parts or faces.

Last few layers in a CNN are typically fully-connected layers, as the ones found in DNNs (described in previous section), taking advantage of strong features generated by the previous layers to determine a class estimation.

3.4 Modality-Specific Architecture in Deep Learning

We refer to this construction comprising of two types of hidden layers – hidden layers related to a specific sensor type and hidden layers that capture unified concepts across sensor types – as Modality-Specific Architecture (MA). In this construction, separate architectures are built for each modality to first learn sensor-specific information before their generated concepts are unified through representations that bridge across all the sensors (i.e., shared modality representations) later in the network (as illustrated in Figure 3). MA is based on the architecture proposed in [52], although our formulation and experiments represent the first time this solution has been tested on mobile sensor data.

Deep Neural Networks (MA-DNNs). Formalizing the earlier high-level description of a multimodal deep learning architecture: the state ($\mathcal{A}_i^{L+1}$) of each individual DNN layer: (x_i^{L+1}) of layer ($L+1$) is dependent on the unit weights connecting the j^{th} node in layer L to the i^{th} node in layer $L+1$. The output is determined by the activation function, for example for a logistic activation function this can be formulated as:

$$\mathcal{A}_i^{L+1} = \frac{1}{1 + \exp(-\sum_j w_{ij}^{L+1} x_j^L)} \tag{1}$$

As shown in Figure 3, separate architectural branches (M_k) exist for each sensor type without any intra-branch connections between layers until later unifying cross sensor layers (U_l) in the larger multimodal architecture. While M_k layers learn representations tied to a single modality (such as the accelerometer), U_l layers seek to

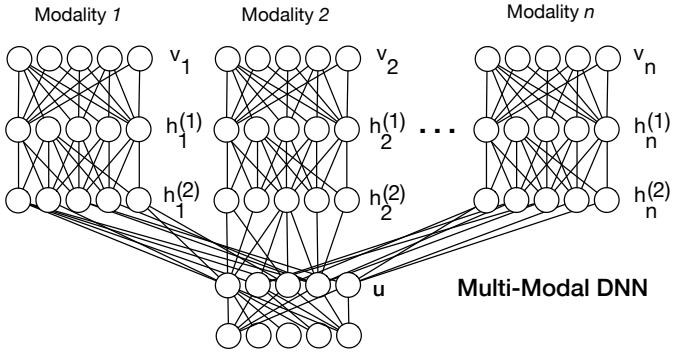


Fig. 3. Modality-specific Deep Neural Networks. The network adopts a split architecture: separate branches exist for each of n modalities, which are then joint in the unifying cross-sensor layers.

learn representations that fuse information between sensors. Collectively, all layers contribute to the learning of a joint representation of all sensor modalities; in other words, $P(\mathbf{v}_{acc.}, \mathbf{v}_{gyro.}, \mathbf{v}_{gps}, \ldots \mid \Theta)$ where Θ spans all model parameters. With respect to the architecture, a key hyper-parameter is the depth (i.e., the number of layers) of every M_k branch and U_l. This changes the complexity and richness of feature representation learned at each architectural branch. Some sensors are simple (such as a light indicator) requiring little representational power, others are complex (such as audio data) and benefit from the rich degree of processing to capture the information they contain. As a result, it is common for M_k to be of variable depth across different branches. The respective depth of each also impacts, at least conceptually, the type of semantic information that is being attempted to be fused between each sensor. Depth is also a factor for U_l in terms of controlling the opportunity for representations that fuse sensors to be discovered, therefore this is impacted by the richness, and sheer number, of the sensors that fan in. As is standard practice, such hyper-parameters are decided with cross-validation at training time.

Multimodal DNN Learning Algorithm. Conventional deep model training processes, like back-propagation with supervised training [14], is well suited for this task, due to the flexibility to split gradients onto each modality branch. In our earlier work [57], we have explored the unsupervised approach as an initial training step with auto-encoders, followed by traditional back-propagation to fine-tune the network feature representations in a construction called Restricted Boltzmann Machines. However, the impact of this extra step on accuracy is not significant enough overall to be deemed important, but the associated increase in training time is a penalty. For that reason we use the traditional back-propagation for the experiments presented in the evaluation section.

Performing Inferences. Post training, inferences with a multimodal deep model occur similarly as with regular DNNs. Sensor data of each type is provided as input to the corresponding architectural branch of the MA classifier. Each branch of the network performs its internal computations independently, same as described for FC before, producing feature signatures influenced by their internal feature representations. These are then combined in the joint part of the network, following a typical forward pass from here.

Convolutional Neural Network (MA-CNN). The multimodal construction using CNNs is analogous to that of MA-DNN. Each sensing modality has its own dedicated CNN to extract preliminary features over several layers (operating as described for FC-CNN). These produce intra-sensor features which are combined through fully-connected layers to identify the target class.

4 EVALUATION

In this section, we empirically compare different techniques for multimodal learning on wearable devices. We evaluate the performance of feature representational learning methods, with two feature concatenation (FC) deep classifiers: FC-DNN and FC-CNN, and two Modality-Specific Architecture (MA) deep classifiers: MA-DNN and MA-CNN, as detailed in Section 3. These are compared against two commonly adopted shallow methods, Random Forest (RF) or Decision Tree (DT), or any task-specific purpose-built technique where available. To offer an overview, our key findings are summaries below:

- Feature representational learning works consistently well across a wide range of activity recognition tasks (recognition of activity, gait, sleep stage and indoor-outdoor), outperforming shallow classifiers throughout while avoiding reliance on hand-tuned dedicated features.
- MA-CNN gives the best accuracy in 3 out of 4 datasets studied despite the nature of classification tasks being very different. Energy consumption measurements indicate this is also sustainable on common wearable devices. This makes MA-CNN a strong candidate as a default classifier for activity recognition and context detection with wearables and mobile devices.
- MA deep classifiers outperform FC on all four datasets, achieving accuracies that are on average 5% better. The difference in accuracies between the MA and FC approaches is most obvious in complex classification tasks, such as activity recognition, where MA outperform by up to 16%. Nevertheless, both MA and FC based deep classifiers achieve better accuracies than shallow classifiers.

4.1 Methods

This subsection details the datasets and baselines used for evaluation.

Datasets. We consider four publicly available datasets in order to represent custom setups in a wide range of activity recognition tasks performed on wearable and mobile devices. Table 1 summarises key features of the data.

STISEN Heterogeneity Activity Recognition Dataset, collected and studied by Stisen et al. [65]. This dataset contains readings of two motion sensors, Accelerometer and Gyroscope, from 9 users performing 6 activities (‘Biking’, ‘Sitting’, ‘Standing’, ‘Walking’, ‘Stair Up’, ‘Stair Down’). Great device diversity is captured in the dataset, each participant collecting data with 8 different smartphones and 4 smartwatches.

GAIT A dataset comprising of data from 460 participants, which forms a diverse sample of the population with distribution across ages (8 to 78 years old) and genders. Previously studied by Ngo et al. [53] with highly engineered signal-based solutions. Gait recognition is done on 5 classes: walking on flat surface, walking up slope, walking down slope, descending stairs and ascending stairs. The data was captured with two inertial sensors (accelerometer and gyroscope), each sampling in triaxial dimension. Remarkable to this dataset is the large population sample and diversity as mentioned earlier.

Sleep-Stage (SS) The Sleep-EDF Database [19], part of PhysioNet, contains physiological data (two EEG readers, one EOG and one EMG) collected from 20 people, annotated with 6 sleep stages (‘Awake’, ‘Stage 1’, ‘Stage 2’, ‘Stage 3’, ‘Stage 4’, ‘REM’). All participants in this dataset suffer from sleep disorders, making it substantially difficult to find simple patterns across all subjects.

Indoor-Outdoor (IO) This dataset [56] contains smartphone sensor readings (light, proximity, magnetic, microphone, cell, battery thermometer), from two different phones and annotated with ‘indoor’ or ‘outdoor’. This was collected in 3 different environments (university campus, city center, residential areas), which brings high variations in the signal patterns as previously highlighted in [56].

Dataset	No. of users	No. of classes	No. of modalities
STISEN	9	6	2
GAIT	460	5	2
Sleep-Stage	20	5	4
Indoor-Outdoor	2	2	7

Table 1. Summary of datasets used in our evaluation. Each dataset is selected based on its intrinsic complexity, such as the number of users (GAIT), diversity of sensing devices (STISEN), diversity of participants (SS) and number of sensing modalities (IO).

Baselines and tools. This section introduces the classifiers used for comparison on the four datasets summarized before (Table 1). We consider the following popular shallow classification techniques as our benchmark:

- Random Forest (RF): shallow-classifier-based feature concatenation, ensemble of decision tree classifiers;
- Decision Tree (DT): shallow-classifier-based feature concatenation; C4.5 is often used in wearable devices due to the low resource footprint and effectiveness in types of activities with few degrees of freedom (small feature space), e.g. walking or running.

For each recognition task, we compare the performance of deep classification techniques (FC-DNN, FC-CNN, MA-DNN, MA-CNN) with shallow techniques (RF and DT). In the case of the GAIT dataset, we also extend the comparison to five other purpose-built inference models, previously considered in the literature and established as best-performing for gait recognition [53].

We compare the performance of these classifiers based on the F1-score, defined as the harmonic mean of precision and recall, $\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$. This metric is sensitive to misclassifications, and also robust to unbalanced distributions of samples across classes.

Methodology. Evaluation of shallow classifiers was performed in line with their original works, such as employing the same features as described in [65], for activity recognition or by extracting ECDF features (which shows good performance in our evaluation despite its compression effect). In cases where an earlier analysis is not available (e.g. for SS), training and testing of traditional classifiers are performed with Weka [23].

Datasets are split into training set and test set following the leave-one-outmethod, permuting which instance is left aside as test data and averaging across all iterations to get the final performance. It is never trivial to train deep neural networks optimally, so to speed up the training process, we performed a random search to identify the most appropriate hyperparameters (including depth and number of nodes) for each task, guided by the best F1-score on the test set.

Deep neural networks are particularly good at extracting their own internal features representation, performing well directly on raw data. The only intervention on these datasets was to normalize the sampling rate with a low-high pass filter, this being common practice when using Android devices due to the irregular sampling rate. Sampling frequency is chosen in alignment with previous analysis on such datasets. Appropriate time window size is another task-specific parameter as well as the overlapping between signal segments (experimenting with values between 50% and 70%), which control the amount of useful information provided to the classifier as independent inputs and increase the number of training samples respectively. We consider these the minimal preprocessing requirements for training with deep neural networks. To highlight the advantages of just minimal preprocessing of data for training deep neural networks, we perform other common preprocessing transformations like frequency domain (by

transforming the time window signal using the Fast Fourier Transform) and the Empirical Cumulative Distribution Function (ECDF) features at specific interest points, to compare with.

4.2 Comparison of Multimodal Context Recognition Techniques

Here we present the results achieved by the previously described multimodal techniques on the four context recognition tasks: activity recognition (Stisen dataset), gait recognition (Gait dataset), sleep stage detection (SS dataset) and indoor vs. outdoor detection (IO dataset).

Figure 4 shows the average F1-scores of selected classifiers. We interpret these results taking the following views:

Deep vs. Shallow Classifier. Deep learning classifiers outperform shallow classifiers across all four datasets, generalizing across diverse tasks without the pain of features identification as required when working with traditional shallow classifiers. When comparing deep solutions against common shallow classifiers (DT, RF), the average accuracy difference is substantial, 27%. CNN-based architectures dominate in all but one dataset (GAIT), where a task-specific approach [53] performs slightly better.

Deep: Feature Concatenation vs. Modality-Specific Architecture. Table 2 presents the F1 scores for best performing MA deep classifiers, FC deep classifiers and shallow classifiers across the four datasets. From this we see that MA consistently outperforms FC deep architectures in terms of accuracy. This is an indicator that early representations on each sensing modality help to discriminate between classes right from the first few layers of the network, in contrast to concatenated modalities inputs which mix data representations too early, thus missing valuable insights within each sensing modality.

Feature Concatenation: Deep vs. Shallow. Considering both shallow and deep classifiers adopting FC, FC-deep classifiers on average outperform FC-shallow classifiers, by 24% in many cases.

	MA- deep	FC- deep	Best-performing shallow
STISEN	**81.6**	70.36	74.5 (RF)
GAIT	89.5	88.6	**93.22** (NGO2014)
Sleep-Stage	**66.4**	65.1	55.04 (RF)
Indoor-Outdoor	**82.3**	80.1	58.92 (RF)

Table 2. Comparison of MA-deep classifier, FC-deep classifier and the best performing shallow classifier for each dataset. We highlight the best-performing architecture under each task.

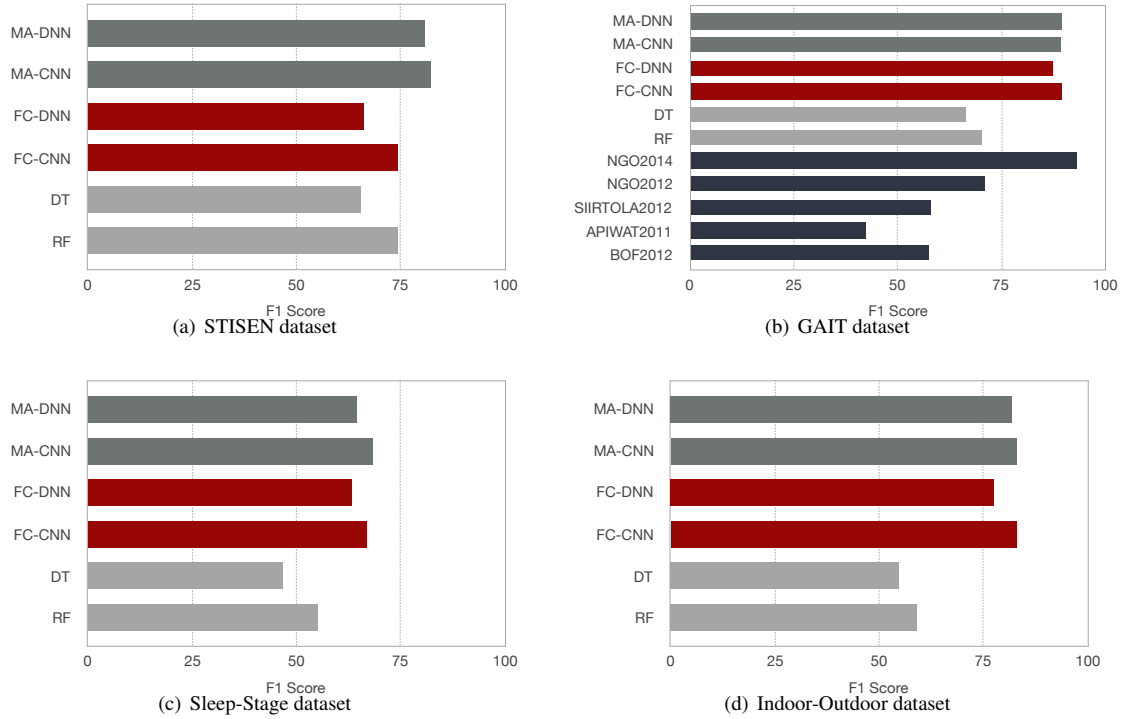


Fig. 4. F1 score achieved by different classifiers, (MA- and FC- neural networks based methods, and shallow or purpose-built methods) on the four datasets. This evaluation is performed with input samples in time domain for deep learning classifiers and features extracted specifically for each dataset to use in shallow classifiers. Despite the effort with shallow classifiers to extract the most relevant features for each domain task (based on previous studies for each dataset), deep learning classifiers still perform better without such requirements.

4.3 Activity recognition with large device diversity: STISEN dataset

The first experiment is conducted on the Stisen dataset, which features data from only 9 users but with large device diversity. To obtain reliable user-independent results, we perform training with the leave-one-out policy, so that data from each subject is used in turn once as test data, while data from all other eight subjects is contained in a training set. In the preprocessing phase, sample rate of sensor data collected with Android devices is normalized to 50Hz, and segmented in time windows of 2 seconds with overlapping of 50%. This is required to guarantee that inputs to neural networks are always the same size and capture a constant time window.

Results for this experiment averaging over 9 subjects are presented in Figure 4(a), which shows that MA deep classifiers achieve the best accuracies, with F1-score of 82%, while FC deep classifier and shallow classifiers both achieving just about 70%-75%.

A more in-depth perspective is provided in Figure 5, where the accuracy of each deep classifier is presented per subject. This shows that MA deep classifiers are able to maintain an accuracy above their FC counterparts. It is clear

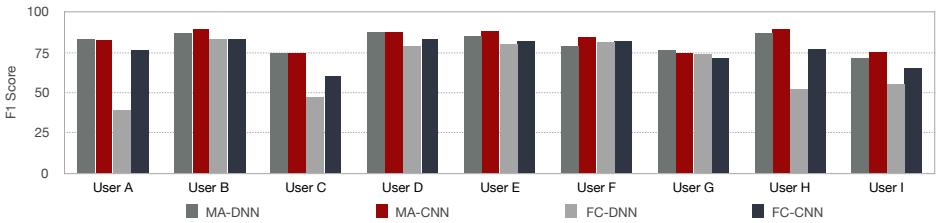


Fig. 5. Per-user comparison of the four deep learning methods MA-DNN, MA-CNN, FC-DNN, and FC-CNN on the activity recognition task using the Stisen dataset.

that FC deep classifiers perform suboptimally for some users, indicating a bad generalization due to not identifying relevant intra-modality feature representations, while the MA is uniformly better across all users.

Though deep learning methods achieve better performance than shallow classifiers, the complexity of this dataset has an impact on the accuracy magnitude, with values below 90%. To improve this, we found that incremental training with the MA-DNN architecture, where less than 5 minutes of labeled data from the same device (customizing the model to one individual-device pair) is enough to boost performance by about 18%, well into the desirable region.

Feature representation learning. Deep classifiers taking raw data (time domain) as input are found to produce better results than shallow classifiers, which must rely on adequate features selection. We consider the following alternative forms of preprocessing transformations on our input data:

- Fast Fourier Transformation (FFT), input data in frequency domain;
- Empirical Cumulative Distribution Function (ECDF) with chosen sample points.

These transformations are akin to features extraction as their role is to filter the raw data stream into a different, more compact representation.

In Figure 6 we observe that, when compared to raw data inputs, applying these transformations in the data preprocessing stage lowers the accuracy of MA deep classifiers (also trained with transformed samples), by almost 5%. This decrease in accuracy of MA classifiers after applying feature extraction (data transformations) suggests that MA deep classifiers are more capable to perform inferences directly from raw data. Access to raw data allows these classifiers to extract their own unfiltered representation of strong features in sensor signals.

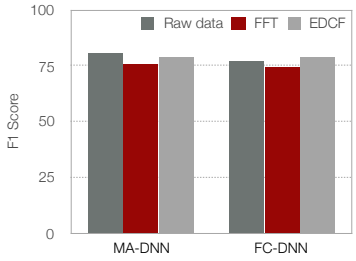


Fig. 6. Comparison of accuracies achieved by classifiers without features extraction (no data transformation) and with features extraction (data transformation with FFT and ECDF) on the STISEN dataset.

4.4 Activity recognition with large number of participants: GAIT dataset

This recognition experiment is performed on the GAIT dataset, which joins data from a very large number of participants (460), with a broad demographic distribution. We split this dataset into two sections, used for training and test, following the same distribution as presented in [53].

In this experiment, we compare deep classifiers with purpose-built solutions and shallow classifiers. As before, we evaluate four deep classifiers: FC-DNN, FC-CNN, MA-DNN and MA-CNN. For benchmarking these, we use the same shallow classifiers (DT and RF) and additionally some earlier purpose-built solutions engineered for this task and dataset alone as presented in [53]: NGO2014, NGO2012, SIIRTOLA2012, APIWAT2011, and BOF2012.

Figure 4(b) presents the performance of above mentioned models, showing that both representational learning methods (FC and MA) and the purpose-built shallow classifier, NGO2014, produce very good results. The highest accuracy was achieved by NGO2014 at 93.2%, followed by MA-DNN and FC-CNN at 89.7%. It is worth noting here that this gap in accuracy between the highly-engineered purpose-built detector, NGO2014, and the general purpose deep classifiers is surprisingly narrow. This indicates that general representational learning methods can closely match the performance of purpose-built methods, potentially limited only by data surplus and training effort.

Figure 7 presents a per action class comparison of performance across the best-considered models. It is easy to observe the consistently good performance of deep classifiers across classes. Further, Figure 8 shows a per user cumulative distribution of the F1 scores achieved by FC and MA deep classifiers. This distribution shows that even on a per-user level, MA deep classifiers outperform FC deep classifiers, with above 90% accuracy in more than 50% of the cases.

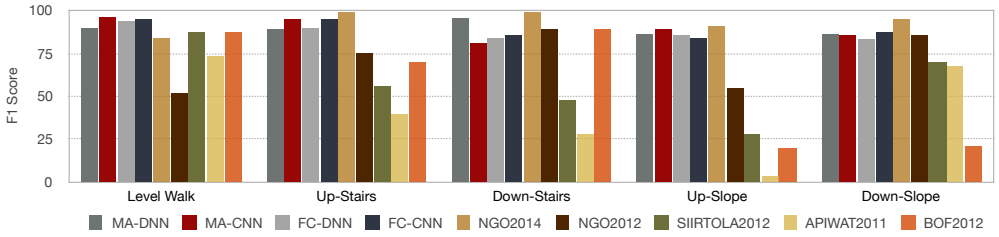


Fig. 7. Per-action class comparison between the performance of all deep learning classifiers and purpose-built solutions on the GAIT dataset.

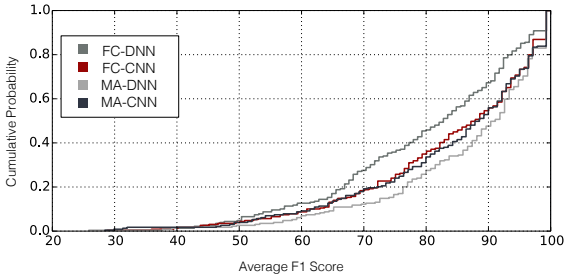


Fig. 8. Cumulative distribution of per-user F1 score of deep classifiers on the GAIT dataset.

Feature representation learning. Here we compare different preprocessing transformations for this dataset, similar to those presented for the Stisen dataset. As shown in Figure 9, data transformations significantly lower the accuracy of all deep classifiers (by 11% on average) compared to inference on raw data. This decrease in accuracy after applying feature extraction confirms earlier observations made on the Stisen dataset, suggesting that deep classifiers are capable of performing inference directly on raw data much better than when these are interpreted by various transformations.

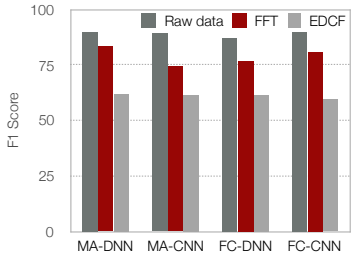


Fig. 9. Comparison of accuracies achieved by deep classifiers without feature extraction and with features extraction (data transformation with FFT and EDCF) on the GAIT dataset.

4.5 Sleep Stage Detection

This experiment considers other sensing modalities, different from the accelerometer and gyroscope available in the previous two datasets, detecting sleep stage from physiological data signals.

The SS dataset has 4 modalities with a sampling frequency of 100Hz. In the preprocessing stage, the sampling frequency is down-sampled to 10Hz to reduce the size of input to our neural networks (and thus the computational costs), while also capturing a large enough time window of 10 seconds for each classification instance. We find that sleep stages 1 and 3 are highly similar, so we group them as a single class. Features for shallow classifiers were extracted using the ECDF method presented before.

From the results presented in Figure 4(c), we can see that feature representational learning methods achieve much better accuracies than shallow methods. On average, we find that deep classifiers enable a 29% accuracy improvement over shallow methods. When looking at a break down of F1-scores per class (Table 3), we find that this improvement is mainly attributed to an almost doubling of performance on classes 'Sleep Stage 4' and 'REM', which shallow classifiers find particularly difficult to detect. MA deep classifiers are slightly more accurate than FC by about 2%, while shallow classifiers are clearly suboptimal for this task. One possible explanation is the simplicity of chosen features to train shallow classifiers, though it also highlights the difficulty of identifying relevant features in new and unexplored detection tasks.

Feature representation learning. Figure 10 presents a comparison between time domain signals input (raw data) and the same two signal transformations as before, FFT and ECDF provided as input to the four deep classifiers. While the performance of deep classifiers on the raw data is nearly 70%, operating on signal transformations reduces the accuracy to about 12% on average.

	MA-DNN	MA-CNN	FC-DNN	FC-CNN	RF	J48 (DT)
Wake	66.13	**70.94**	65.06	69.28	51.28	39.74
Stage 1 & 3	27.59	**30.03**	25.07	29.50	33.15	33.63
Stage 2	77.08	**80.18**	76.19	79.65	70.62	60.43
Stage 4	68.40	**79.59**	68.40	74.20	44.53	31.28
REM	72.48	**75.91**	70.86	73.94	47.33	37.09
Weighted average	64.50	**68.22**	63.19	67.01	55.04	46.59

Table 3. F1 scores for all classifiers in each sleep stage detection on the Sleep-Stage dataset.

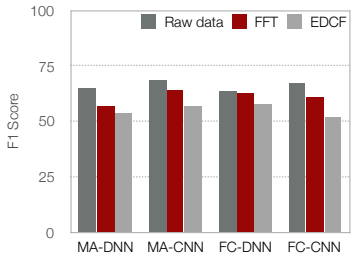


Fig. 10. Comparison of accuracies achieved by classifiers without features extraction and with features extraction (data transformation FFT and ECDF) on the Sleep-Stage dataset.

4.6 Indoor-Outdoor Detection

The IO dataset has its own unique characteristics which make this an interesting exploration – 7 independent and different sensing modalities. It is a binary classification task, though highly diverse across the three environments where data was collected from.

'Indoor' and 'outdoor' areas are sampled from three environments: Campus, City Centre, Residential area, which we labeled as 'env1', 'env2' and 'env3' respectively. We perform the training with the leave-one-out (environment) method, so that each environment takes turn in being the test dataset while the other two assist in training the classifier. Table 4 presents results of each classifier over the three iterations.

Figure 4(d) again confirms that feature representational learning methods are successful, as they achieve much better accuracy than shallow methods (by 43% on average). This observation still holds on a per-environment level, as shown in Table 4. It can be observed that deep classifiers perform consistently better across all environments. Another observation is that all classifiers perform significantly poorer when tested with 'env2', suggesting that it is the hardest environment to detect; this is also true taking a geographical perspective of the areas where data was collected from, with the university Campus and the Residential environments being in closer geographical proximity and farther away from the City Centre. We also see that within shallow classifiers, Random Forest perform better than Decision Tree by 7%.

Feature representational learning. In Figure 11 and as previously observed, classification on raw data with deep classifiers achieves the best performance, here about 81%. After preprocessing the input data with previously mentioned transformations (FFT and ECDF), the accuracy of classifiers drops by about 13% on average.

Training set	Test set	MA-DNN	MA-CNN	FC-DNN	FC-CNN	DT	RF
env2 + env3	env1	87.75	**87.87**	84.5	**87.87**	77.05	68.45
env1 + env3	env2	65.44	**67.38**	57.62	**67.38**	26.6	38.7
env1 + env2	env3	92.64	**92.97**	90.19	**92.97**	60.95	69.63

Table 4. F1 scores for cross-environment evaluation on the Indoor-Outdoor dataset.

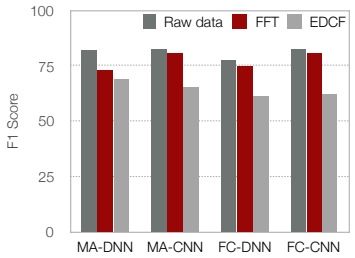


Fig. 11. Comparison of accuracies achieved by deep classifiers without feature extraction and with features extraction (data transformation with FFT and ECDF) on the Indoor-Outdoor dataset.

5 MOBILE HARDWARE FEASIBILITY

Deep architectures exert significant resource challenges on embedded platforms, mainly due to their high demands of memory, computations and energy. In the following, we present runtime experiments on two embedded platforms, namely Snapdragon 400 and Snapdragon 800 SoCs, see Figure 12. Feasibility experiments presented in this section are focused mainly on system resource-usage of the deep architectures. Deployment experiments are performed with an efficient hand-tuned implementation of all deep models presented earlier in this work.

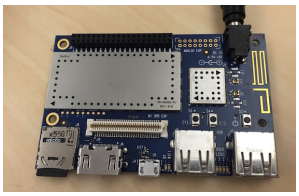
5.1 Target Mobile Hardware

Qualcomm Snapdragon 400. While targeting wearable devices, this Qualcomm processor offers similar performance to many smartphones. It is found within a range of smartwatches, such as the LG G Watch R [42] and includes a quad-core 1.4 GHz CPU and 1 GB of RAM. Additional GPU and DSP processors are also available, but due to a lack of driver support we are forced to use the CPU only for experiments.

Qualcomm Snapdragon 800. As the second embedded platform we use the Snapdragon 800 SoC and run all the deployment experiments on this platform and measure energy consumption and overall runtimes. As in the case with Snapdragon 400, we only use the CPU on Snapdragon 800 to execute the deep models.

5.2 Model Runtime Implementation

To assess the resources demanded by the various multimodal DNNs validated in the prior section, a shared runtime is implemented that executes the inference stage (only) of each model. This prototype is realized through a mix of modules from the Torch [67] ported individually to each processor, that is supported by a set of custom C/C++ components implemented by the authors. Although Torch introduces a degree of overhead, as it acts as an interpreter for the high-level Lua language (in which we encode all models and their parameters), it also offers a number of low-level, highly optimized mathematical operation APIs, which are useful for deep model executions. That said,



(a) Snapdragon 400

(b) Snapdragon 800

Fig. 12. The development board for profiled wearable-class hardware; processors and measured energy profiles of the boards are identical to that of processors found in commercial wearables. For example, the Snapdragon 400 (a) is found within smartwatches, such as the LG G Watch R Smartwatch and (b) Snapdragon 800 is found within Samsung Galaxy 9005 and Nokia Lumia 1520.

certain native Torch operations are replaced with C/C++ extensions to exploit processor-specific opportunities for execution speedup, and better memory management. Furthermore, additional components (e.g., FFT library) are used for any conventional feature extraction as needed by the model specification.

Overall, this implementation can be characterized as adopting best practices understood by those who regularly hand-optimize deep learning models – either for scalability, or in this case operation in resource-limited environments. More obvious examples of incorporated optimizations include keeping of minimal model architectures needed by the inference stage, and the profiling of the data flow of runtime execution to understand memory and cache bottlenecks; or even changing the power profile of processor components to improve the trade-off between execution times and energy use.

5.3 System Resource Usage Experiments

For each of the four activity datasets evaluated in the prior section, we examine runtime resource demands, e.g., memory, computation and energy, of all deep models studied in this paper. Performance comparisons are drawn across all models on two embedded platforms.

The memory requirements of different types of deep model architectures trained on individual data sets are summarized in Table 5. While storing individual model parameters we use 32-bit precision. The largest model (33.1 MB) was found to be the FC-DNN, trained on the STISEN dataset and the smallest model (0.2 MB) was the MA-DNN model trained on the Indoor-Outdoor dataset.

Table 6 and 7 respectively illustrates the average running time (in milli-seconds) and energy consumption (in mJ) of the deep models observed on the Snapdragon 400 platform. We repeated each inference 1000 times and took the average time and energy to mitigate the effect of inherent variations in OS executions. Similarly in Table 8 and 9 we present the runtime results as observed on the Snapdragon 800 platform, which is heavily impacted by the Android scheduling system, resulting in poorer performance than Linux based Snapdragon 400. Despite this, results indicate that the deep models can be executed efficiently on embedded platforms.

Lastly, in Figure 13, we present a variation of CPU load and memory requirement observed on the Snapdragon 400 platform, while executing a MA-CNN model trained on the Gait dataset. The MA-CNN model begins by executing two convolutional layers in parallel and keeping the CPU load almost 100%. The convolution layers require small number of parameters and this keep the overall memory demand low. Once the convolution operations are completed, the CPU load drops to a lower value as the OS becomes occupied in loading the parameters from memory, making memory demand rise. Once all the parameters are loaded in the memory the CPU load becomes rises again to obtain the final inference result.

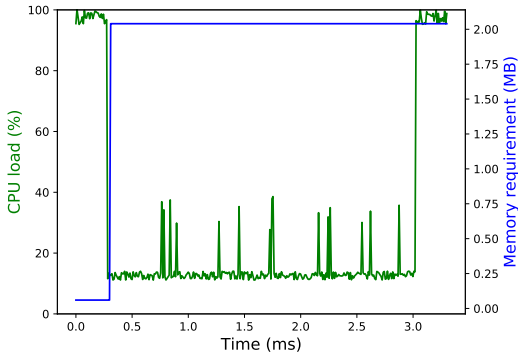


Fig. 13. CPU load and memory requirement against time.

	STISEN (MB)	GAIT (MB)	Sleep-Stage (MB)	Indoor-Outdoor (MB)
MA-DNN	22.1	12.5	1.9	0.2
MA-CNN	8.4	18.6	3.7	0.4
FC-DNN	33.1	2.1	0.3	0.6
FC-CNN	8.4	6.0	0.3	0.6

Table 5. Trained Model sizes (in MBytes) across all data sets under 32-bit precision.

	STISEN (ms)	GAIT (ms)	Sleep-Stage (ms)	Indoor-Outdoor (ms)
MA-DNN	2.8	1.9	1.5	1.7
MA-CNN	9.7	3.3	2.5	0.9
FC-DNN	3.5	2.2	0.8	0.5
FC-CNN	6.5	7.7	1.5	0.7

Table 6. Average model execution time (milli-seconds) observed on Snapdragon 400.

	STISEN (mJ)	GAIT (mJ)	Sleep-Stage (mJ)	Indoor-Outdoor (mJ)
MA-DNN	5.4	3.7	2.9	3.3
MA-CNN	18.7	6.4	4.8	1.7
FC-DNN	6.7	4.2	1.5	1.0
FC-CNN	12.5	14.8	2.9	1.3

Table 7. Average energy consumption (milli-Joule) due to individual deep models observed on Snapdragon 400.

	STISEN (ms)	GAIT (ms)	Sleep-Stage (ms)	Indoor-Outdoor (ms)
MA-DNN	38.8	26.3	20.8	23.6
MA-CNN	134.4	45.7	34.7	12.5
FC-DNN	48.5	30.5	11.0	6.9
FC-CNN	90.1	106.7	20.8	9.7

Table 8. Average model execution time (milli-seconds) observed on Snapdragon 800.

	STISEN (mJ)	GAIT (mJ)	Sleep-Stage (mJ)	Indoor-Outdoor (mJ)
MA-DNN	63.6	43.2	34.1	38.6
MA-CNN	220.5	75.0	56.8	20.5
FC-DNN	79.6	50.0	18.2	11.4
FC-CNN	90.1	106.7	34.1	15.9

Table 9. Average energy consumption (milli-Joule) due to individual deep models observed on Snapdragon 800.

6 DISCUSSION AND LIMITATIONS

This section discusses the practicality of using deep learning for general ubiquitous computing tasks, observed from our experience, describing the difficulties encountered and choices made in training deep neural networks; continuing the discussion with current limitations and the opportunity for future work.

Sensor Signal Preprocessing. From the experiments reported in Section 4, it is clear that operating directly on raw data (i.e., time domain), instead of preprocessing data with FFT or ECDF is not only more computationally efficient (avoiding data transformations) but also effective in training a more accurate model. Choosing an optimal time window size is also important and very specific to the detection task. When activity classes are very similar, a larger time window can help to capture more information – the obvious downside is that a larger input increases computation cost for the neural network. We find that it is a good practice to start with small time windows (such as 30 ms) and gradually expand to capture more information.

The Training Process. A good strategy in training is to start with a small network (such as two layers with a small number of neurons per layer) and gradually expand the size and complexity of the network driven by observations on a separate set of instances (validation set).

When calibrating the network, we consider the key factors: bias, variance, training data distribution. High bias occurs when the model is performing poorly on both training and validation datasets; this is usually addressed by increasing the size of the model or that of the training set. High variance occurs when the classifier overfits training data with good performance on training set but poor on validation set, in which case reducing the network size or introducing regularization (discussed below) can help. In situations where datasets are too small, alternative solutions have proven successful in eliminating this limitation by adopting Active Learning in Bayesian DNNs [16].

In the following we detail some assumptions and decisions made in this exploration:

Learning Rate. In the traditional stochastic gradient descent, this indicates the weights update strength in back-propagation. Careful selection of its value is important: a large value may never converge, while a value too small may take infinitely long to converge. A typical value is 1e-3, though no two training sets are the same, so variations need to be considered to achieve a good convergence. In our exploration, values between 1e-1 and 1e-5 were the

most effective. Learning rate also works in pair with momentum which encourages updates when gradients are consistently in the same direction. Adam algorithm for learning rate policy was determined to be very efficient in many cases.

Number of epochs. The training process executes forward and backward propagations over the entire dataset in one epoch. It is typical for networks to converge very slowly on a complex dataset, affected by noise in data or high similarity in classes. In this situation, a higher number of epochs is required to converge, lengthening the training time. With too many epochs the network may just produce insignificant updates, which should be detected by an early stop policy when the network has converged. In our implementation, we restricted the number of epochs to a maximum of 400 with an adaptive learning rate policy.

Network initialization. The simplest approach to initialize the weights between neurons and biases is with random small values chosen from a mean zero and one variance distribution. Auto-encoders are another solution in initializing the network before supervised learning. From unlabelled data instances an auto-encoders can extract features by reproducing the input to the output on sections of the network. However, in our case datasets are already labeled and networks have only a small number of layers so the impact of auto-encoders is minimal.

Dropout ratio. The dropout layer is a common method for regularization due to its simplicity and efficiency. This implies randomly dropping connections between neurons during training to avoid reliance on single paths through the network (dominant neurons). Its disadvantage comes from extending training time due to more combinations of network connections needing to be reinforced for the network to learn effectively. Training time is affected by dropout factor and width of the connecting layers.

Batch Normalization. This is another very efficient solution for regularization. With the presence of a Batch Normalization layer, weights are normalized again after each update, which allows for a larger value for learning rate to be used – faster training time.

We make our implementation code available as a training framework dedicated to multimodal sensing data for other researchers to use in their work [18].

Additional Deep Learning Methods. Advances in deep learning continue to proliferate and produce a growing range of potential avenues with the potential to improving modeling of wearable and mobile sensor data. It is important to note that in this investigation we concentrated only on performing inferences on static frames, discarding their temporal connection with other frames. This can be seen as one shot classification, not requiring to track sensor signals for a long period of time, ideal for applications requiring occasional sensing. However, other solutions like Recurrent Neural Networks (RNNs) [20] and Long Short-Term Memory (LSTM) can take advantage of this time correlation to improve performance even further. We leave this as open opportunity for further research.

7 RELATED WORK

Applications of Multimodal Learning. Multimodal learning has a vast application domain. Applications have been seen in audio-visual speech recognition [52], image captioning [63], machine translation [34], sentiment analysis [55] and affect recognition [30]. In the space of ubiquitous computing, example applications include human activity recognition [1], sleep detection [12] and emotion recognition [36]. Many recognition tasks were previously only primarily performed with unimodal learning, with the availability of low-energy sensors, many such tasks are recently explored using multimodal learning. For example, authentication models involve both gaze and touch recognition [32], or eating recognition might involve motion of head, wrist and audio [47].

Multimodal Sensor Fusion. Conceptually, classification models based on multimodal sensor data have a clear relationship to techniques of sensor fusion. Within sensor networks, and more broadly in fields such as robotics,

fusion methods routinely leverage different sensor types for purposes like localization [6, 10, 54, 61]. However, the fusion techniques developed are difficult to directly apply to the design of learning algorithms and the feature representations they operate on. More direct insights and methods are found in the design of classifiers of related domains, such as: computer vision, scene understanding [21] and affective computing [30] – as well as numerous examples of existing models of context and activity recognition. In short: the use of, and benefiting from, the input of multiple sensor types is nothing new. But understanding how to combine significantly different sensor inputs is still an open problem. Although multimodal models are routinely seen, the tools we have to construct them (ranging from ensembles of separate classifiers to co-training or simply collapsing data types into single feature vectors) each have their shortcomings; and even simple questions (such as, at what stage should data types be merged?) must still be addressed on a case by case basis (e.g., [62]).

Multimodal Deep Learning. A prime example in the general space of multimodal deep learning is audio-visual speech recognition [50], where much work has been done using neural networks [52]. A number of neural networks have been proposed to perform multimodal deep learning, including CNN [51], RBM [52] and RNN [46]. The choice of neural network often depends on the type of recognition involved, as there is currently no consensus on which network would work best. For instance, in tasks where sequential data is involved (e.g. image sentence description [46]), multimodal versions of recurrent neural networks have been frequently proposed to handle these tasks. While there is work comparing a small number of multimodal learning methods, such as [8] which compares decision tree classifiers with back propagation neural networks, we note that there has not been a comprehensive case study comparing a greater number of deep and shallow multimodal learning architectures. Finally, we wish to highlight an early version of this study was presented in poster form [57] – though the work presented here is of course a significant extension.

Deep Learning in Ubiquitous Computing. Only recently has the exploration into deep learning methods for mobile sensing scenarios begun (e.g., [24, 39]). But with the diversity of exploration rapidly expanding [3, 7, 16, 17, 25, 31, 40, 49, 71]. To the best of our knowledge, the work presented here is the first time that the detection of indoor/outdoor context and transportation mode has been attempted with any form of deep learning, even for single sensor modalities. There is still much to be understood in how such models should be architected, and which variety of algorithms will be most effective – our work adds to this knowledge, that is still in a nascent stage. Closely related models to those we propose in this work are found in [52]. However, we use simpler *Restricted* Boltzmann Machines rather than the *deep* version described in [52] (although, both are still forms of deep learning). Similarly, [33, 64] concern themselves with multimodal models but focus tightly on learning features. None of these papers consider mobile sensor data types nor the classification objectives we study here. Furthermore, few consider a mobile platform as the operating environment of their models. In fact, little multimodal study of this aspect of our work exists, although broad understanding of resource-limited deep learning is accelerating [4, 11, 26, 29, 37, 38, 59, 68] and we expect many existing results to extend to multimodal formulations, though this still remains to be verified.

8 CONCLUSION

In this paper, we perform a systematic study of multimodal deep learning architectures to assess how and when these new techniques satisfy the exigencies of activity and context inferences with mobile devices. We present experiments with four distinct variants of deep neural networks across very diverse and difficult context detection datasets, while comparing their performance with common shallow classifiers and hand-crafted task-specific detectors. Two of these variants are state-of-the-art in deep learning architectures for performing modalities fusion used in other scenarios (video, voice, text) – here, referring to as MA-DNN and MA-CNN – and are for the first time used with wearables and mobile sensing devices for activity recognition and context detection. Experiments that

span a wide range of sensor types, competing multimodal learning algorithms, and activity and context detection tasks, collectively show our proposed general-purpose deep approach to multimodal sensor fusion modeling is both broadly applicable and is able to exceed the performance of previous general solutions and even match task-specific sensor-tuned solutions. This innovation in sensor data modeling is complemented with a practical proof-of-concept implementation designed to measure the overhead of these techniques on two state-of-the-art mobile/wearable processors. Results show that devices that adopt the deep modeling approach, emphasized here, are able to maintain sustainable norms of size, weight and lifetime despite the increased complexity of deep learning methods.

ACKNOWLEDGMENTS

This project received funding from the European Commission's Horizon 2020 research and innovation programme under grant agreement No 687698, through a HiPEAC Collaboration Grant. We thank all the anonymous reviewers for their constructive comments, which helped us to improve the quality of this work.

REFERENCES

[1] Michael Barz, Mohammad Mehdi Moniri, Markus Weber, and Daniel Sonntag. 2016. Multimodal Multisensor Activity Annotation Tool. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. ACM, New York, NY, USA, 17–20. https://doi.org/10.1145/2968219.2971459

[2] Yoshua Bengio, Ian J. Goodfellow, and Aaron Courville. 2015. Deep Learning. (2015). http://www.iro.umontreal.ca/~bengioy/dlbook Book in preparation for MIT Press.

[3] S. Bhattacharya and Nicholas D. Lane. 2016. From smart to deep: Robust activity recognition on smartwatches using deep learning. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. 1–6. https://doi.org/10.1109/PERCOMW.2016.7457169

[4] Sourav Bhattacharya and Nicholas D. Lane. 2016. Sparsification and separation of deep learning layers for constrained resource inference on wearables. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*. ACM, 176–189.

[5] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[6] Tatiana Bokareva, Wen Hu, Salil Kanhere, Branko Ristic, Neil Gordon, Travis Bessell, Mark Rutten, and Sanjay Jha. 2006. Wireless sensor networks for battlefield surveillance. In *Proceedings of the land warfare conference*. 1–8.

[7] Heike Brock, Yuji Ohgi, and James Lee. 2017. Learning to judge like a human: convolutional networks for classification of ski jumping errors. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*. ACM, 106–113.

[8] Donald E. Brown, Vincent Corruble, and Clarence Louis Pittard. 1993. A comparison of decision tree classifiers with backpropagation neural networks for multimodal classification problems. *Pattern Recognition* 26, 6 (1993), 953 – 961. https://doi.org/10.1016/0031-3203(93)90060-A

[9] Andreas Bulling, Jamie A. Ward, and Hans Gellersen. 2012. Multimodal recognition of reading activity in transit using body-worn sensors. *TAP* 9, 1 (2012), 2. https://doi.org/10.1145/2134203.2134205

[10] Jose A Castellanos and Juan D Tardos. 2000. *Mobile robot localization and map building: A multisensor fusion approach*. Kluwer academic publishers.

[11] Guoguo Chen, Carolina Parada, and Georg Heigold. 2014. Small-footprint Keyword Spotting Using Deep Neural Networks. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'14)*.

[12] W. Chen, A. Sano, D. L. Martinez, S. Taylor, A. W. McHill, A. J. K. Phillips, L. Barger, E. B. Klerman, and R. W. Picard. 2017. Multimodal ambulatory sleep detection. In *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*. 465–468. https://doi.org/10.1109/BHI.2017.7897306

[13] Tanzeem Choudhury, Gaetano Borriello, Sunny Consolvo, Dirk Haehnel, Beverly Harrison, Bruce Hemingway, Jeffrey Hightower, Predrag "Pedja" Klasnja, Karl Koscher, Anthony LaMarca, James A. Landay, Louis LeGrand, Jonathan Lester, Ali Rahimi, Adam Rea, and Danny Wyatt. 2008. The Mobile Sensing Platform: An Embedded Activity Recognition System. *IEEE Pervasive Computing* 7, 2 (April 2008), 32–41. https://doi.org/10.1109/MPRV.2008.39

[14] Li Deng and Dong Yu. 2014. *DEEP LEARNING: Methods and Applications*. Technical Report MSR-TR-2014-21. http://research.microsoft.com/apps/pubs/default.aspx?id=209355

[15] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Çağlar Gülçehre, Vincent Michalski, Kishore Reddy Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, Raul Chandias Ferrari, Mehdi Mirza, David Warde-Farley, Aaron

Courville, Pascal Vincent, Roland Memisevic, Christopher Pal, and Yoshua Bengio. 2015. EmoNets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces* (2015), 1–13. https://doi.org/10.1007/s12193-015-0195-2

[16] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian Active Learning with Image Data. *CoRR* abs/1703.02910 (2017). http://arxiv.org/abs/1703.02910

[17] Petko Georgiev, Sourav Bhattacharya, Nicholas D. Lane, and Cecilia Mascolo. 2017. Low-resource Multi-task Audio Sensing for Mobile and Embedded Devices via Shared Deep Neural Network Representations. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 50 (Sept. 2017), 19 pages. https://doi.org/10.1145/3131895

[18] Github repository 2017. Multimodal Deep Learning Framework. https://github.com/vradu10/deepfusion.git. (2017).

[19] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* 101, 23 (2000), e215–e220. Circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215 PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[20] Alex Graves, A-R Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 6645–6649.

[21] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. 2010. Multimodal semi-supervised learning for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 902–909.

[22] Haodong Guo, Ling Chen, Liangying Peng, and Gencai Chen. 2016. Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble. In *Proceedings of UbiComp*. ACM.

[23] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* 11, 1 (Nov. 2009), 10–18. https://doi.org/10.1145/1656274.1656278

[24] Nils Hammerla, James Fisher, Peter Andras, Lynn Rochester, Richard Walker, and Thomas Plötz. 2015. PD Disease State Assessment in Naturalistic Environments using Deep Learning. In *AAAI 2015*.

[25] Nils Hammerla, Shane Halloran, and Thomas Ploetz. 2016. Deep, Convolutional, and Recurrent Models for Human Activity Recognition using Wearables. In *Proceedings of IJCAI*. ACM.

[26] Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* (2015).

[27] Awni Y. Hannun, Carl Case, Jared Casper, Bryan C. Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep Speech: Scaling up end-to-end speech recognition. *CoRR* abs/1412.5567 (2014). http://arxiv.org/abs/1412.5567

[28] Samuli Hemminki, Petteri Nurmi, and Sasu Tarkoma. 2013. Accelerometer-based Transportation Mode Detection on Smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems (SenSys '13)*. ACM, New York, NY, USA, Article 13, 14 pages. https://doi.org/10.1145/2517351.2517367

[29] Loc N Huynh, Youngki Lee, and Rajesh Krishna Balan. 2017. DeepMon: Mobile GPU-based Deep Learning Framework for Continuous Vision Applications. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 82–95.

[30] Ashish Kapoor and Rosalind W Picard. 2005. Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 677–682.

[31] Thomas Kautz, Benjamin H Groh, Julius Hannink, Ulf Jensen, Holger Strubberg, and Bjoern M Eskofier. 2017. Activity recognition in beach volleyball using a Deep Convolutional Neural Network. *Data Mining and Knowledge Discovery* (2017), 1–28.

[32] Mohamed Khamis, Florian Alt, Mariam Hassib, Emanuel von Zezschwitz, Regina Hasholzner, and Andreas Bulling. 2016. GazeTouchPass: Multimodal Authentication Using Gaze and Touch on Mobile Devices. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 2156–2164. https://doi.org/10.1145/2851581.2892314

[33] Yelin Kim, Honglak Lee, and E.M. Provost. 2013. Deep learning for robust feature generation in audiovisual emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. 3687–3691. https://doi.org/10.1109/ICASSP.2013.6638346

[34] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *CoRR* abs/1411.2539 (2014). http://arxiv.org/abs/1411.2539

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[36] Saewon Kye, Junhyung Moon, Juneil Lee, Inho Choi, Dongmi Cheon, and Kyoungwoo Lee. 2017. Multimodal Data Collection Framework for Mental Stress Monitoring. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers (UbiComp '17)*. ACM, New York, NY, USA, 822–829. https://doi.org/10.1145/3123024.3125616

[37] Nicholas D. Lane, S. Bhattacharya, P. Georgiev, C. Forlivesi, L. Jiao, L. Qendro, and F. Kawsar. 2016. DeepX: A Software Accelerator for Low-Power Deep Learning Inference on Mobile Devices. In *2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. 1–12. https://doi.org/10.1109/IPSN.2016.7460664

[38] Nicholas D. Lane, Sourav Bhattacharya, Petko Georgiev, Claudio Forlivesi, and Fahim Kawsar. 2015. An early resource characterization of deep learning on wearables, smartphones and internet-of-things devices. In *Proceedings of the 2015 International Workshop on Internet of Things towards Applications*. ACM, 7–12.

[39] Nicholas D. Lane and Petko Georgiev. 2015. Can Deep Learning Revolutionize Mobile Sensing?. In *HotMobile 2015*.

[40] Nicholas D. Lane, Petko Georgiev, and Lorena Qendro. 2015. DeepEar: Robust Smartphone Audio Sensing in Unconstrained Acoustic Environments Using Deep Learning. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 283–294. https://doi.org/10.1145/2750858.2804262

[41] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep Learning. *Nature* (2015).

[42] LG G Watch R 2017. LG G Watch R. https://www.qualcomm.com/products/snapdragon/wearables/lg-g-watch-r. (2017).

[43] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. 2016. Multimodal Emotion Recognition Using Multimodal Deep Learning. *CoRR* abs/1602.08225 (2016). http://arxiv.org/abs/1602.08225

[44] Hong Lu, Jun Yang, Zhigang Liu, Nicholas D. Lane, Tanzeem Choudhury, and Andrew T. Campbell. 2010. The Jigsaw Continuous Sensing Engine for Mobile Phone Applications. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems (SenSys '10)*. ACM, New York, NY, USA, 71–84. https://doi.org/10.1145/1869983.1869992

[45] Lumo Lift 2017. Lumo Lift. http://www.lumobodytech.com. (2017).

[46] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. 2014. Explain Images with Multimodal Recurrent Neural Networks. *ArXiv e-prints* (Oct. 2014). arXiv:cs.CV/1410.1090

[47] Christopher Merck, Christina Maher, Mark Mirtchouk, Min Zheng, Yuxiao Huang, and Samantha Kleinberg. 2016. Multimodality Sensing for Eating Recognition. In *Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '16)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, 130–137. http://dl.acm.org/citation.cfm?id=3021319.3021339

[48] Microsoft Band 2017. Microsoft Band. http://www.microsoft.com/Microsoft-Band/. (2017).

[49] Francisco Javier Ordóñez Morales and Daniel Roggen. 2016. Deep convolutional feature transfer across mobile activity recognition domains, sensor modalities and locations. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers*. ACM, 92–99.

[50] Y. Mroueh, E. Marcheret, and V. Goel. 2015. Deep multimodal learning for Audio-Visual Speech Recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2130–2134. https://doi.org/10.1109/ICASSP.2015.7178347

[51] Sebastian Münzner, Philip Schmidt, Attila Reiss, Michael Hanselmann, Rainer Stiefelhagen, and Robert Dürichen. 2017. CNN-based Sensor Fusion Techniques for Multimodal Human Activity Recognition. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers (ISWC '17)*. ACM, New York, NY, USA, 158–165. https://doi.org/10.1145/3123021.3123046

[52] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, Lise Getoor and Tobias Scheffer (Eds.). Omnipress, 689–696.

[53] Trung Thanh Ngo, Yasushi Makihara, Hajime Nagahara, Yasuhiro Mukaigawa, and Yasushi Yagi. 2015. Similar gait action recognition using an inertial sensor. *Pattern Recognition* 48, 4 (2015), 1289 – 1301. https://doi.org/10.1016/j.patcog.2014.10.012

[54] Reza Olfati-Saber and Jeff S Shamma. 2005. Consensus filters for sensor networks and distributed sensor fusion. In *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC'05. 44th IEEE Conference on*. IEEE, 6698–6703.

[55] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174, Part A (2016), 50 – 59. https://doi.org/10.1016/j.neucom.2015.01.095

[56] Valentin Radu, Panagiota Katsikouli, Rik Sarkar, and Mahesh K. Marina. 2014. A Semi-supervised Learning Approach for Robust Indoor-outdoor Detection with Smartphones. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems (SenSys '14)*. ACM, New York, NY, USA, 280–294. https://doi.org/10.1145/2668332.2668347

[57] Valentin Radu, Nicholas D. Lane, Sourav Bhattacharya, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. 2016. Towards multimodal deep learning for activity recognition on mobile devices. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 185–188.

[58] Valentin Radu and Mahesh K. Marina. 2013. HiMLoc: Indoor Smartphone Localization via Activity Aware Pedestrian Dead Reckoning with Selective Crowdsourced WiFi Fingerprinting. In *In Proc. Indoor Positioning and Indoor Navigation (IPIN)*. IEEE. http://dx.doi.org/10.1109/IPIN.2013.6817916

[59] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. 2016. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision*. Springer, 525–542.

[60] Devendra Singh Sachan, Umesh Tekwani, and Amit Sethi. 2013. Sports Video Classification from Multimodal Information Using Deep Neural Networks. In *2013 AAAI Fall Symposium Series*.

[61] Gyula Simon, Miklós Maróti, Ákos Lédeczi, György Balogh, Branislav Kusy, András Nádas, Gábor Pap, János Sallai, and Ken Frampton. 2004. Sensor network-based countersniper system. In *Proceedings of the 2nd international conference on Embedded networked sensor systems*. ACM, 1–12.

[62] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. 2005. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 399–402.

[63] Kihyuk Sohn, Wenling Shang, and Honglak Lee. 2014. Improved Multimodal Deep Learning with Variation of Information. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 2141–2149. http://papers.nips.cc/paper/5279-improved-multimodal-deep-learning-with-variation-of-information

[64] Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal Learning with Deep Boltzmann Machines. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger (Eds.). Curran Associates, Inc., 2222–2230. http://papers.nips.cc/paper/4683-multimodal-learning-with-deep-boltzmann-machines.pdf

[65] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart Devices are Different: Assessing and Mitigating Mobile Sensing Heterogeneities for Activity Recognition. In *The 13th ACM Conference on Embedded Networked Sensor Systems*.

[66] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[67] Torch 2017. Torch. http://torch.ch/. (2017).

[68] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*. IEEE, 4052–4056. https://doi.org/10.1109/ICASSP.2014.6854363

[69] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. 1083–1092.

[70] Pengcheng Wu, Steven C.H. Hoi, Hao Xia, Peilin Zhao, Dayong Wang, and Chunyan Miao. 2013. Online Multimodal Deep Similarity Learning with Application to Image Retrieval. In *Proceedings of the 21st ACM International Conference on Multimedia (MM '13)*. ACM, New York, NY, USA, 153–162. https://doi.org/10.1145/2502081.2502112

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

[72] Piero Zappi, Thomas Stiefmeier, Elisabetta Farella, Daniel Roggen, Luca Benini, and Gerhard Tröster. 2007. Activity Recognition from On-Body Sensors by Classifier Fusion: Sensor Scalability and Robustness. In *3rd Int. Conf. on Intelligent Sensors, Sensor Networks, and Information Processing (ISSNIP)*. 281–286. http://www2.ife.ee.ethz.ch/~droggen/publications/wear/EDAS_ISSNIP.pdf

Received February 2017; revised August 2017; accepted October 2017

Multimodal Deep Learning for Activity and Context Recognition

Valentin Radu[†], Catherine Tong[§], Sourav Bhattacharya[‡], Nicholas D. Lane[‡§],
Cecilia Mascolo[*], Mahesh K. Marina[†], and Fahim Kawsar[‡△]

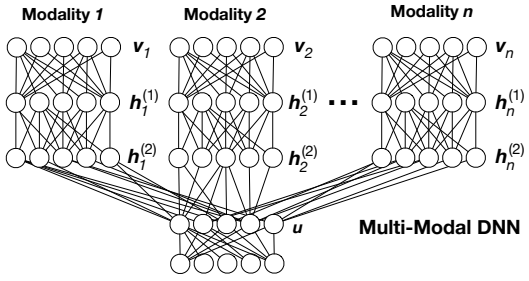
[†]University of Edinburgh, [§]University of Oxford, [‡]Nokia Bell Labs,
[*]University of Cambridge, [△]TU Delft

The popularity of smart mobile and wearable devices has given rise to a growing interest in complex sensing tasks such as activity and context recognition. These tasks typically rely on data from a multitude of modalities, captured by low-energy small form-factor sensors such as light detectors, magnetometer, accelerometer and barometer. A successful combination of information across multi-modal sensor streams dictates the fidelity at which they can track user behavior and context.

In this study, we consider the benefits of adopting *deep learning* algorithms for activity and context recognition as captured by multi-sensor systems. Specifically, we use fully-connected Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) and compare two multi-modal architectures for each type of neural network. One architecture, *Feature Concatenation* (FC), is a commonly employed approach for multimodal data fusion, which simply concatenate raw sensor streams at the input layer. We compare this to a novel architecture, *Modality-Specific Architecture* (MA); In this architecture (Fig. 1), separate neural networks are built per modality, before their generated concepts are unified through representations which bridge across all sensors. MA is based on the architecture proposed in [Ngiam et al., 2011], although our formation and experiments is the first time that this architecture has been tested on mobile data [Radu et al., 2018].

We use 4 publicly available datasets for evaluation, covering recognition of human activity, gait, sleep stages, as well as indoor-outdoor detection. Our experiments show that these generic multimodal neural network models compete well with shallow methods and task-specific modelling pipelines, across a wide range of sensor types and inference tasks. Although the training and inference overhead of these multimodal deep approaches is in some cases appreciable, we also demonstrate the feasibility of on-device mobile and wearable execution is not a barrier to adoption. This study is carefully constructed to focus on multimodal aspects of wearable data modeling for deep learning by proving a wide range of empirical observations, which we expect to have considerable value in the community. We summarize our observations into a series of practitioner rules-of-thumb and lessons learned that can guide the usage of multimodal deep learning for activity and context detection.

Figure 1: Modality-Specific Architecutre (MA) with Deep Neural Networks. Separate branches exist for each of n modalities, which are joined in unifying cross-sensor layers.



References

[Ngiam et al., 2011] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In Getoor, L. and Scheffer, T., editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 689–696. Omnipress.

[Radu et al., 2018] Radu, V., Tong, C., Bhattacharya, S., Lane, N. D., Mascolo, C., Marina, M. K., and Kawsar, F. (2018). Multimodal deep learning for activity and context recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(4):157:1–157:27.

Appendix C

Conference Paper: Pervasive Health 2019

This work was co-supervised by Nic Lane and Gabriella Harari, a psychologist from Stanford University. We are currently revising the manuscript for submission to the International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2019).

Predicting Big-Five Personality Using Large-scale Networked Mobile and Appliance Data

We present the first large-scale (9110-user) study of data from both mobile and networked appliances for Big-Five personality inference. Building on methods (viz. features and classifiers) previously successful for personality detection from mobile-*only* data, this investigation shows Big-Five personality can be detected with accuracies of 77% (similar levels as other studies) under this setting – despite the size and complexity (mix of mobile and appliance) of the dataset. This result acts as a replication study of techniques that are commonly utilized in the literature, but under a type of dataset previously never studied. Moreover, we offer ancillary results, in particular, as to behavior and physical health features that correlate with mobile and appliance data and how inference accuracy alters as cohort scale and diversity change. Collectively these results provide initial insights as to how to model data from mobile and appliances for use-cases likely beyond personality inference, for instance, wellbeing and mental health. We anticipate our findings are timely given the rapid uptake by consumers for smart devices in the home, which will cause datasets of this type to be more readily available in the near future.

CCS Concepts: • **Human-centered computing** → *Ubiquitous and mobile devices*; • **Applied computing** → *Psychology*;

ACM Reference format:
. 2018. Predicting Big-Five Personality Using Large-scale Networked Mobile and Appliance Data. 1, 1, Article 1 (May 2018), 11 pages.
https://doi.org/0000001.0000001

1 INTRODUCTION

Personality traits describe a person's characteristic patterns of thinking, feeling, and behaving. From an applied perspective, knowing a user's personality could be useful for device customization (e.g. personalization based on psychological characteristics) and for understanding contributing factors to their wellbeing. Traditionally, personality traits are measured using self-report surveys, which is time-consuming and can be difficult to scale up in commercial settings. However, mobile sensing technologies permit unobtrusive collection of real-world behavioral patterns [11]. Such technologies may be used to classify personality traits passively, without requiring methods such as experience sampling [12], or other related survey instruments.

A number of prior studies have been conducted looking into the automated inference of personality [3–5, 8, 9, 15, 17, 21], and other psychological states (e.g. mood [16, 18], stress [2]) from everyday digital technologies. The results from past studies suggest that people's personality traits manifest in ways that can be captured by measurements from digital media devices (e.g., smartphone sensors and phone logs) and platforms (e.g., social media). Moreover, features from these measurements can predict a person's self-rating on the Big Five personality traits (Extraversion, Agreeableness, Conscientiousness, Emotional-stability and Openness). However, past studies tend to (i) be small-scaled and focused on homogeneous samples (under 200 people), (ii) collect data only from one digital media source (e.g. smartphone), (iii) typically in experimental settings, and (iv) focuses primarily on measurements of sociability and mobility behavior (e.g. call logs, location information).

XXXX-XXXX/2018/5-ART1
https://doi.org/0000001.0000001

Given these limitations, our primary question is: whether existing approaches for personality inference be applied to data which (i) involves a large, diversified user population, (ii) collects data from a wide mixture of digital media sources (including mobile devices and home networked appliances) which could contain sparse measurements, (iii) uses real-world data that are naturally collected from these devices in-the-wild as people use them, and (iv) focuses on other types of measurements about health and physiological behaviors. Increasingly data with such properties is becoming readily available as consumers integrate a mixture of *smart*-devices (e.g., bathroom scales, in-bed sleep trackers, along with the more common phones and wearables) into their lives. This in turn raises the importance of understanding the opportunities for performing societal-scale personality and wellbeing experiments this data will afford, that typically are done with mobile-*only* data [16, 18].

To answer this question, in this study, we examine the accuracy of inference when predicting the Big Five personality traits of users based on a dataset with the aforementioned characteristics. Our dataset consists of measurements from 9110 users of Withings devices [1] (also known as Nokia Digital Health), collected from the range of devices owned by the user (ranging from smart watches to blood pressure monitors). To the best of our knowledge, this is the first large-scale study reporting on a dataset with a mixture of networked mobile and appliance data sources for personality inferences.

We deliberately design this study to build on past research methods and make our results comparable to existing work. As such, we follow the general inference approach adopted in [3, 4, 23, 24], that proved successful at a smaller scale on mobile-centric data. Our work begins with a replication study of such methods with our Withings dataset. Results show that comparable accuracy levels for personality inference, as prior work, are indeed possible using the data. We then examine the individual merits of using data sourced from different digital media devices: only appliances (i.e., weighing scale, sleep tracker, blood pressure monitor), only mobile devices (i.e., smartwatch), and a mixture of appliances and mobile devices (i.e., smartwatch, weighing scale, sleep tracker and blood pressure monitor). Next, we evaluate the scalability of personality inference models by assessing their performance across a spectrum scale of users, from 83 users (same as that in [3]) to 9110 users. Finally, we summarise and discuss our findings, providing insights and lesson learned for future studies.

2 BACKGROUND

In this section, we provide background relating to personality studies in psychology research and survey existing automated approaches for personality inference.

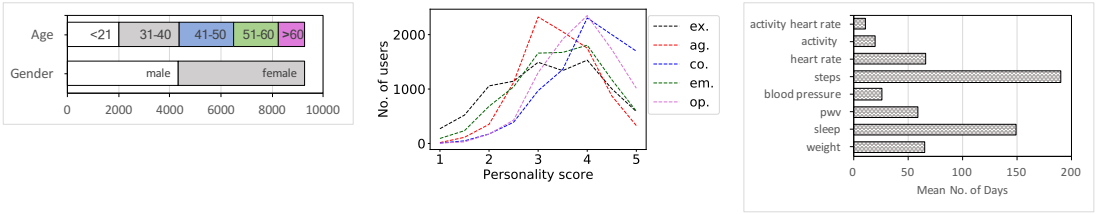
The Study of Personality Personality traits describe a person's characteristic patterns of thinking, feeling, and behaving. The most widely used model for measuring personality focuses on the Big-Five personality traits - Extraversion, Agreeableness, Conscientiousness, Emotional-stability and Openness to experiences. Personality inventories, consisting of adjectives to describe the respondent, are common instruments for assessing the personality dimensions. The 44-item Big Five Inventory [13] and the Ten-Item Personality Inventory (TIPI) [10] have both been employed in automated personality inference research, for instance in [5, 15] and [3, 4] respectively.

There have been several attempts to predict personality traits by modelling data from everyday digital media technologies. Broadly, the past work can be described as taking two approaches: (1) a platform-based approach that focuses on data from various social media platforms, and (2) a device-based approach that focuses on data from various mobile devices.

Platform-based approach Existing work in personality inference using data from social media primarily focus on two kinds of information: linguistic features and social network information. Linguistics features have been shown to be related to personality traits, [8, 9] applied linguistic analysis onto tweets and Facebook profile information to predict traits. [8, 9, 15] have also used social network information such as friendships, in personality inference; while[14] have shown that personality traits could be predicted from Facebook likes. Such

Fig. 1. Descriptive information of dataset.

(a) Demographics. (b) Distribution of personality trait scores. (c) Mean number of days per modality.



studies have demonstrated the viability of personality prediction from social media platforms. Thus, we focus our efforts on device-based approaches in the present research.

Device-based approach More related to our study are device-based approach to personality inference [3–5, 17]. [17] uses low-level sensor data from sociometric badges to study individual and group behaviour, using high-level behavioural descriptions such as physical and speech activity. Although the goal of [17] is not personality inference, their results show that it is possible to correlate high-level behavioural information with personality traits. [3–5, 21] all utilise data from smartphones for personalty inference, using information such as call logs, use of Bluetooth. [21] focuses on inference through building a picture of the social network of smartphone users, while [5] proposes behavioural indicators for inference, e.g. regularity and diversity found in calls/ texts, spatial behavior from GPS signals. In [3, 4] the authors use aggregated features from mobile logs and sensor data (e.g. call/SMS logs, BT logs) for personality inference, we find their studies to be the most applicable on our dataset for direct comparison, therefore we use their approach and results as guidance and reference throughout this paper.

More generally, studies that examine personality inferences from physiological data is lacking so far. Part of the reason for this lack of research may be the need for additional device hardware which might be a prohibiting factor in collecting such measurements. However, there have been prototypes developed for recognising emotions from physiological signals such as skin temperature and ECG signals [7]. It has been suggested that the advantages of using such data include the fact that physiological signals may offer a more direct view into our psychological states and may be 'difficult to fake' [16]. Thus, in the present research we explore physiological measurements and their relationship to personality by focusing on physiological measurements from the following types of devices: weighing scale, sleep tracker, and blood pressure monitor.

3 METHODOLOGY

In this section, we discuss the dataset and study design.

3.1 Dataset

Our dataset comprises of personality scores and behavioral and physiological data of 9270 users of *Withings* devices. The user population spans a diverse demographic spectrum, as seen in Figure 1(a).

The TIPI Scale. Individuals included in the study are voluntary respondents of an online questionnaire sent out in March 2017. Respondents complete the TIPI, a 10-item survey instrument used for assessing Big-Five personality dimensions, as part of a larger questionnaire. Following the survey design [10], final scores for each personality dimension are computed, resulting in scores ranging from 1 to 5 in intervals of 0.5, where a higher score means a stronger association with the personality trait. Figure 1(b) shows the distribution of personality scores for the entire dataset. In our analysis, we split each trait into high-scoring and low-scoring groups through the median for binary classification. After removing users who had invalid responses and unrealiable sensor measurements, our dataset consists of 9110 individuals.

Table 1. Dataset Sensor Measurements.

Modality	Measurements	Source
Steps	step count, gait speed	smart watch (mobile)
Sleep	sleep duration, bed-in/out times, time to sleep, time to wake, no. of times of being awake, awake duration, duration of awake/ light/ REM†/ deep sleep,	smart watch (mobile), sleep tracker (appliance)
Weight	weight, BMI.	scale (appliance)
Heart rate	average/ minimum/ maxmimum heart rate	smart watch (mobile), sleep tracker, scale (appliance)
Pulse wave velocity (pwv)	pwv pwvH (a score indicating the healthiness of pwv)	scale (appliance)
Blood pressure (sbp)	sbp	blood pressure monitor (appliance)
Activity	activity duration, calories burned, activity heart rate	smart watch (mobile)

†Only available through appliance.

Behavioral and Physiological Data. Table 1 lists all measurements taken from the following *Withings* devices.

- *Phone app.* A repository for measurements taken from all Withings devices for user's monitoring and goal-setting purposes. Manual entry for missing data is also available (e.g. user may enter weight/ height.)
- *Smart watch.* Wearable device which comes in different models, provides inference data on activity, sleep (and heart rate in certain models).
- *Sleep tracker.* WIFI-enabled pad placed under mattress, provides inference data on sleep and heart rate.
- *Weighing scale.* WIFI-enabled scale which comes in different models, provides raw weight measurements (as well as inference data on pulse wave velocity and heart rate in certain models).
- *Blood pressure monitor.* WIFI-enabled device, provides raw data on blood pressure and heart rate.

Each user's data from all devices that he/she owns are collected in daily resolution over a time window of up to 1 year. The amount of data present per user is dependent on usage, and Figure 1(c) shows the mean number of days per modality per person in the dataset.

Raw and Inference Data. Like many studies in the area, our models build on lower-level activity inference models that are fairly mature in the industry; in [23, 24], the authors have relied on inferences such as walking, running, sleep stages build on prior classifiers. Our inference data is based on commercial-quality models by Withings, and we have verified through discussions that they have validated their models in similar fashions to those paper in the literature. (i.e., controlled user trials) at a similar scale. In addition, these devices are used by millions of people on a daily basis - an added check since other methods have not been tested against. Further, we highlight that our dataset, being a mixture of raw information (e.g. weight, blood pressure) as well as inference data, represents an increasingly popular form of data useful for wellbeing research.

Unique Dataset Characteristics. Our dataset is different from other commonly studied datasets for personality inference in the following manners:

- *Large-scale, diversed population.* There is no pre-selection of the subjects for this study, any Withings device user who has voluntarily completed the survey is included.
- *Data collection in-the-wild.* Individuals collect data as they use these devices in their everyday lives.
- *Sparsity.* Amount of data collected is dependent on user's usage and ownership of different devices.

- *Mix of data sources.* Data are collected from both mobile devices and home networked appliances.
- *Mix of raw and inference data.* Data consists of raw measurements as well as behavioral data inferred with commercial-quality models.
- *Focus on behavioral and physiological features.* Primary functions of Withings devices relate to health and wellbeing, so the dataset focuses on other types of measurements about health and physiological behaviors

3.2 Method

This study aims to verify that existing methods can be applied onto our dataset for personality inference, which is an application domain without strong baselines for direct comparisons. Therefore, we reference and adopt the general approach taken in many related works looking at sensor-based datasets for inference of wellbeing information [3–5, 23]. Such general approach is a 2-step process of first looking for features, then applying a machine learning model for inference. In detail, the approach first considers a measurement time series of a suitable window length, then obtaining features by aggregating the time series, followed by feature selection with correlation-based analysis or established feature selection algorithms, and finally using a machine learning model is used to infer the interested trait or well-being state.

Benchmark. In terms of the inference problem, the most closely related to our work amongst these related works are [3, 4], as they also considered a binary classification problem for Big-Five personality traits. The dataset considered in [3] (which we refer to as `Phone83users`) consists of smartphone data from 83 users over 8 months, with logs of calls, SMS, Bluetooth scans, application usage; while the dataset in [4] (which we refer to as `Phone117users`) consists of smartphone data from 117 users over 17 months, with logs of calls, SMS, Bluetooth scans, calling profiles and application usage. We use the accuracies achieved in their work as benchmarks.

4 RESULTS

In this section, we present results from a replication study of existing methods with our Withings dataset, which shows comparable accuracies with prior work at 77%. We then present experiments examining the merits of using different data sources, as well as evaluating the scalability of inference models.

4.1 Replication Study

In this subsection, we apply the general approach taken in existing work onto our dataset.

Motivation. This serves as a first step in establishing that personality inference is feasible with this new type of data and features. Through correlation analysis we also hope to gain more insights into the relationships present between modalities and traits.

Setup. To extract features, we split time series of measurements across months, aggregating events by taking mean, minimum, maximum, standard deviation, kurtosis, variability and count. We follow a basic feature selection procedure based on pairwise Pearson correlation test between each feature and each trait. This is followed by feature selection, where we rank the features by their correlation strengths and impose a cut-off of $|r| > 0.05$, and eliminate features which contain too few examples as well as those which may be redundant until we have the number of features below 20. Having selected the features, we impose some dataset *inclusion requirements*: (i) user-months must have least 7 measurements for each modality, (ii) users must have at least 30 measurements in total for each modality. Having collected a dataset of valid users, we consider a binary classification of each trait, using the median to group scores into two classes. We report accuracies achieved using leave-one-cross-validation with SVM and Decision Tree, and a 5-fold cross validation with DNN.

Results. *Features.* We find a number of strong and significant correlations between features and traits ($|r|>0.1$ and $p<0.001$). Many of the correlations found to align with expectations and results from other studies; for

Table 2. List of 17 selected features and their strongest trait correlations found.

Modality	Feature	Strongest correlations	r	No. of items	No. of users
Pulse wave velocity (pwv)	mean pwvH	Ag.	-0.18	2373	702
	mean pwv	Em.	+0.12		
Sleep	mean bed-in	Co.	-0.17	45799	6888
	mean bed-out	Co.	-0.14		
	variability in sleep duration	Co.	-0.10		
	standard dev. in bed-in	Co.	-0.09		
	minimum deep sleep duration	Co.	+0.09		
	standard dev. in light sleep duration	Op.	+0.09		
	standard dev. in bed-out	Co.	-0.09		
	variability in time to sleep	Co.	-0.05		
Blood pressure (sbp)	mean sbp	Ag.	+0.14	1718	408
	no. of sbp measurements	Ex.	-0.10		
Weight	maxmimum weight	Ag.	-0.12	13707	2650
	mean BMI	Co.	-0.08		
	no. of weight measurements	Op.	-0.06		
Heart rate	variability in average heart rate	Co.	-0.06	13630	2270
Steps	variability in gait speed	Ex.	+0.06	31454	5405

Table 3. Left: Accuracy (%) of classifiers as well as baseline (always-majority-class) using 17 behavioural features in the our study. Right: Accuracy(%) reported in `Phone83users` and `Phone117users`.

	Decision Tree	SVM	DNN	Baseline
Ex.	72.1	69.7	**77.5**	66.2
Ag.	79.6	78.9	**84.2**	78.6
Co.	75.6	71.2	**77.1**	71.6
Em.	68.1	63.1	**74.4**	54.4
Op.	**73.6**	63.3	73.4	61.6

	Phone83users	Phone117users
Ex.	75.9	77
Ag.	69.6	75
Co.	74.4	75
Em.	71.5	71
Op.	69.3	74

example, we find a number of sleep features correlating strongly with Conscientiousness; this is not surprising given that [19] has shown that morningness correlates with conscientiousness. We also find that mean BMI correlates negatively with Conscientiousness, which aligns with the relations found in [22]. Through ranking the features by their correlation coefficients, eliminating features which do not have at least 500 user-months, and removing redundant features which may carry overlapping information manually (e.g. mean BMI and mean weight), we arrive at 17 input features with $|r|>0.05$ and $p<0.001$, as listed in Table 2. After enforcing the inclusion criteria, the resulting dataset consists of 168 users with number of examples (N) = 451.

Classificiation. Table 3 presents accuracies achieved by Decision Tree, SVM, DNN, a baseline classifier which always chooses the majority class, as well as benchmarks from [3, 4].

Implications. Application onto our dataset produces results that are consistent with other existing work, even with such different characteristics. Further investigations can be made to enhance feature selection and modelling schemes, but it appears that it is feasible to do personality inference using new types of dataset such as this.

4.2 Modelling smart appliances

The next set of results considers: how does inference performance vary when modelling data from different sources?

Motivation. Modelling data from multiple modalities and sources often present many challenges. However, the effect of using data from different sources on inference performance is not clear as previous studies only considered single channels. To the best of our knowledge, our dataset is the first which enables this comparison to be done for personality inference and we seek to shed light on the matter.

Setup. We introduce three scenarios covering features from different sources: (i) mixture of mobile and appliances, (ii) mobile only, (iii) networked appliance only. In each scenario, the model takes in 12 to 14 features that had been found to be the most strongly correlated to personality traits from each source. Table 4 lists the considered modalities for each scenario.

As before, we consider the users with the complete set of modalities considered. As there are only 586 users of sleep trackers, the inclusion criteria means that the dataset for appliance-only is small compared to the others. To control for scales of the data, we consider this same number of examples (221) in each scenario and compare the inference performance. This binary classification was performed with Decision Tree and DNN.

Table 4. Datasets and considered modalities

Data source	Modalities	No. of users	No. of examples
Mixed source	pwv, sleep (from both mobile and appliance), sbp, weight	168	451
Appliance only	pwv, sleep (appliance only), sbp, weight	61	221
Mobile only	sleep (mobile only), activity, heart rate, steps	715	1377

Results. In Table 5 we compare the accuracies achieved by Decision Tree and DNN across the 3 scenarios. The results of this experiment can be summarized as follows.

With both classifiers, classification using mobile-only data achieves poor performance, particularly for DNN with an average of 36.6% only. This poor performance might be explained by the weak correlations found between mobile-sourced features (other than sleep) with personality traits. Given that there had previously been many studies suggesting activity as correlates of personality traits [20], it is surprising that including it as a feature should incur such low accuracies. We believe this might be a consequence of the sparsity in user's activity data (as seen in Figure 1(c)) which made this modality not suitable for personality inference in our dataset.

In all but one case, appliance-only data provides better performance than alternatives, outperforming mixed-source by an average of 19.0% and 8.5% respectively for Decision Tree and DNN. While this is a first sign that appliance-only data can be a promising alternative to other datasets, we note that the worse performance of mixed-source data might be simply due to noise presented from mobile-sourced data, which seems to be poor personality predictors using data at this scale and type.

Finally, we observe that mixed-source data could be handled better with DNN than Decision Tree. DNN provides the only example (Agreeableness) where a mixed-source classification is better than appliance-only; also, the performance of DNN with mixed-source is within 10% of that of appliance-only, despite the appaling performance of DNN with mobile-only data. This suggests that DNN might be more suitable in handling and dissecting mixed-sourced data to extract useful information.

Implications. Using appliance-only data at a small scale, using behavioral and physiological data achieves good performance, this demonstrates the possibility of using this type of data for wellbeing analysis. However, we also caution that the mobile-only dataset here uses very different features from what had been studied in related studies. It is also only possible to perform the experiment on a small scale, as the number of users who regularly use networked appliances is still a limiting factor.

Table 5. Accuracy (%) of decision tree classifier using features from different sources. Datasets from mixed-source and mobile only were scaled down to have (N=221) for fair comparison with the dataset with appliance only.

	Decision Tree			DNN		
	Mixed source	Appliance only	Mobile only	Mixed source	Appliance only	Mobile only
Ex.	68.4	**87.3**	49.8	77.1	**84.2**	39.5
Ag.	72.9	**81.0**	67.4	**83.7**	82.6	19.8
Co.	67.9	**79.2**	58.8	77.7	**80.8**	30.9
Em.	68.3	**81.4**	53.8	70.5	**88.1**	44.7
Op.	68.8	**83.3**	54.3	73.9	**79.9**	48.1

Table 6. Scale of the considered dataset as modality density criteria are gradually removed.

	Modality density requirement	No. of users	No. of examples
1	pwv, sleep, sbp, weight, heart rate, steps	164	458
2	pwv, sleep, sbp, weight, heart rate	242	647
3	pwv, sleep	1147	3417
4	pwv	1466	4154
5	-	9110	71629

4.3 Scalability analysis

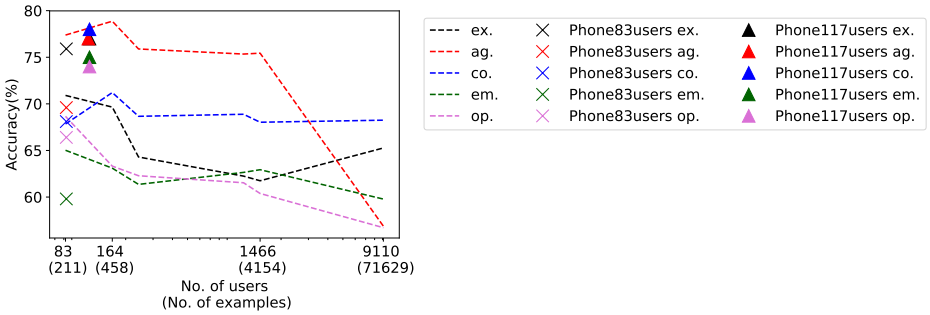
In this subsection, we look into the interplay between the choice of modalities, dataset size, and model performance.

Motivation. In our analysis so far, we see that the dataset scale is greatly diminished when we require users to be active users of all modalities involved in the input features. Since our dataset is of a much greater scale and potentially more complex to model than prior work, this experiment allows us to study the tradeoff between performance and scale, and to gain insights about whether current methods are sufficient for larger scale analysis. To study this, we look at accuracy for this type of new data (appliance and mobile mixture) while holding constant at comparable levels of population size. In addition, we also want to study the impact of demographics on performance as this information is withheld from the studied models.

Setup. We study the inference problem at 6 different scales, from only 83 users to the entire dataset of 9110 users. We consider the 17-item feature list (with 6 modalities, ref. Table 2). To increase the number of users considered, we gradually relax the inclusion criteria such that users must only possess enough data for $(6 - n)$ modalities. Table 6 illustrates this increasing dataset scale and the modality density requirement at each step. At each step, we incrementally train an SVM with data of the newly included users, but letting the model ignore missing modalities. In order to provide a useful reference to compare with `Phone83users`, we also added an additional step where we use the full inclusion criteria but restrict user number to 83 users only.

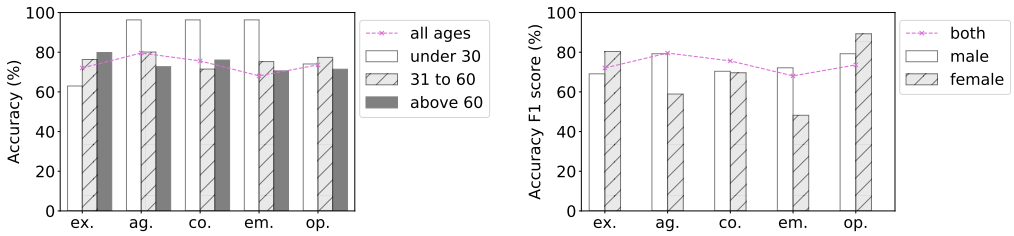
Results. *Scalability.* It should be expected that, at larger dataset scales and with sparser inputs, the performance of the model will be lower (as a consequence of poorer generalization amongst the diverse population and fewer useful information). The results demonstrate a downward general trend on average for the personality traits, matching with this expectation. However, although the performance for Conscientiousness and Extraversion have both gone down at scale, they remain at a reasonable level of accuracy of 68% and 65% respectively even when considering the entire dataset of 9110 users. Conscientiousness has been found to demonstrate strong correlations with most selected features, which might explain its good performance at scale.
Demographic difference. Figure 3 presents accuracy results achieved by the initial Decision Tree model (described in Section 4.1) but separated by user's age and gender groups. It appears that personality classification for under-30s achieves better accuracies on average by 15%, and that for the male population as well by 6.8%. These

Fig. 2. Evaluation results for accuracy as modality requirements are gradually relaxed to increase dataset size.



differences in accuracies might be related to nuanced differences in the way personality manifests in behaviour across age and gender groups; for instance, literature in the psychology domains suggests that significant gender differences exists for emotional stability [25]. The fact that gender and age were held as hidden factors to the classification models might have contributed to these observed differences.

Fig. 3. Evaluation results for different demographic subsets: varying across age groups (left) and genders (right).



Implications. Inference performance using mixed-data sources seems to be scalable to 9110 users for Conscientiousness and Extraversion. However, for the rest (and especially for Agreeableness), accuracy drops rapidly once the dataset size surpasses a thousand users. Better generalization might be achieved if demographic factors are provided to the model.

5 DISCUSSION

Our findings show that using a large-scale dataset with sparsely collected passive behavioural and physiological data is a feasible approach for personality inference. Using existing approach on this dataset, we were able to achieve accuracies in the same range as results in prior work. Due to the mixed-source data available, we are able to present the first evaluation of sourcing data from mobile device or networked appliance. While in this experiment we observe that a poor performance of using mobile-only data, we do not claim that this is the general case for all mobile data since our features for mobile-data are particularly sparse. It remains as our future work to better compare data sources when a suitable dataset becomes available. In our scalability analysis, the observation that the general trend for inference performance goes down as scale increases aligns with our expectations. However, scalability seems to be different for each personality trait. In particular, accuracies for

conscientiousness and extraversion remain at 68% and 65% respectively even when considering the entire dataset of 9110 valid users.

As our work serves to be the first investigation into personality inference using such dataset, we discuss a number of limitations in our work and insights that we believe would be useful for future work:

Personality-dependent inference. From our results, we see that for the same task, variation in accuracies amongst personality traits can be as much as 59%. Rather than using the same features and classifier models for all Big-Five personalities, better performance might be achieved if each prediction problem is tailored to each personality dimension, but not all five at a time.

Regression as inference. Our current work uses the median value to separate users into 2 classes: high-scoring and low-scoring group for each personality trait. We note that while this might be logical in a machine learning context, personality traits are viewed as continuous variables in psychology instead of discrete opposites such as extraversion versus introversion. It is proposed that personalities are to be conceived of as density distributions [6]. Further, we also find that model performance can be sensitive to which group the median value is placed. Given these, it will be worthwhile investigating personality inference as a regression problem in future work.

Demographic influence. Our results regarding demographic variations suggests that the current model (which does not take demographic meta-data as input features) may not be able to generalise the inference performance across population. Results from [25] have suggested that there could be significant differences between personality scores obtained between gender groups. Future work could look into including demographic variables as input features.

Data Sparsity. Although there are 9110 valid users in our dataset, many users are not included in the majority of the analysis in this paper under the current approach as each considered user must have enough data for every input modality. Our future work seeks to examine ways of feature selection and discovery such that the analysed user population size could be maximised.

Lastly, we wish to note that our results seek to demonstrate the possibilities of using appliance-only data for personality inference, which is promising because this might open up a future research space where current models developed for mobile technologies (such as [16]) could be deployed on appliance data.

6 CONCLUSION

In conclusion, we presented the first large-scale (9110-user) study of data from both mobile and networked appliances for Big-Five personality inference. We demonstrated, to varying degrees, that it is feasible to perform personality inference using existing methods with such data, despite its complexity. We also offer ancillary results from investigations of how inference performance is changed using different device sources and at varying data scales. Our findings and insights drawn can help towards performing analysis on this increasingly popular form of dataset in the near-future when people have a ready access to a rich stream of sensor data from mixed sources.

REFERENCES

[1] [n. d.]. Withings. https://health.nokia.com/uk/en/.

[2] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex (Sandy) Pentland. 2014. Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits. In *Proceedings of the 22Nd ACM International Conference on Multimedia (MM '14)*. ACM, New York, NY, USA, 477–486. https://doi.org/10.1145/2647868.2654933

[3] G. Chittaranjan, J. Blom, and D. Gatica-Perez. 2011. Who's Who with Big-Five: Analyzing and Classifying Personality Traits with Smartphones. In *2011 15th Annual International Symposium on Wearable Computers*. 29–36. https://doi.org/10.1109/ISWC.2011.29

[4] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. 2013. Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing* 17, 3 (01 Mar 2013), 433–450. https://doi.org/10.1007/s00779-011-0490-1

[5] Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, and Alex (Sandy) Pentland. 2013. Predicting Personality Using Novel Mobile Phone-Based Metrics. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, Ariel M. Greenberg, William G. Kennedy,

and Nathan D. Bos (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 48–55.
[6] William Fleeson. 2001. Toward a Structure- and Process-Integrated View of Personality: Traits as Density Distributions of States. 80 (07 2001), 1011–27.
[7] A. Gluhak, M. Presser, L. Zhu, S. Esfandiyari, and S. Kupschick. 2007. Towards Mood Based Mobile Services and Applications. In *Proceedings of the 2Nd European Conference on Smart Sensing and Context (EuroSSC'07)*. Springer-Verlag, Berlin, Heidelberg, 159–174. http://dl.acm.org/citation.cfm?id=1775377.1775390
[8] J. Golbeck, C. Robles, M. Edmondson, and K. Turner. 2011. Predicting Personality from Twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. 149–156. https://doi.org/10.1109/PASSAT/SocialCom.2011.33
[9] Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting Personality with Social Media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11)*. ACM, New York, NY, USA, 253–262. https://doi.org/10.1145/1979742.1979614
[10] Samuel Gosling et al. 2003. A Very Brief Measure of the Big-Five Personality Domains. 37 (12 2003), 504–528.
[11] Gabriella M. Harari, Nicholas D. Lane, Rui Wang, Benjamin S. Crosier, Andrew T. Campbell, and Samuel D. Gosling. 2016. Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges. *Perspectives on Psychological Science* 11, 6 (2016), 838–854. https://doi.org/10.1177/1745691616650285 PMID: 27899727.
[12] Joel M Hektner, Jennifer A Schmidt, and Mihaly Csikszentmihalyi. 2007. *Experience sampling method: Measuring the quality of everyday life*. Sage.
[13] O. P. John, Donahue, E. M., and R. L. Kentle. 1991. The Big Five Inventory–Versions 4a and 54. In *Berkeley, CA: University of California,Berkeley, Institute of Personality and Social Research.*
[14] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5802–5805. https://doi.org/10.1073/pnas.1218772110 arXiv:http://www.pnas.org/content/110/15/5802.full.pdf
[15] Lin Li, Ang Li, Bibo Hao, Zengda Guan, and Tingshao Zhu. 2014. Predicting Active Users' Personality Based on Micro-Blogging Behaviors. *PLOS ONE* 9, 1 (01 2014), 1–11. https://doi.org/10.1371/journal.pone.0084997
[16] Robert LiKamWa, Yunxin Liu, Nicholas D. Lane, and Lin Zhong. 2013. MoodScope: Building a Mood Sensor from Smartphone Usage Patterns. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '13)*. ACM, New York, NY, USA, 389–402. https://doi.org/10.1145/2462456.2464449
[17] Daniel Olguín Olguín, Peter A. Gloor, and Alex Pentland. 2009. Capturing Individual and Group Behavior with Wearable Sensors. In *AAAI Spring Symposium: Human Behavior Modeling*.
[18] Kiran K. Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J. Rentfrow, Chris Longworth, and Andrius Aucinas. 2010. EmotionSense: A Mobile Phones Based Adaptive Platform for Experimental Social Psychology Research. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp '10)*. ACM, New York, NY, USA, 281–290. https://doi.org/10.1145/1864349.1864393
[19] Christoph Randler. 2008. MorningnessâĂŞeveningness, sleepâĂŞwake variables and big five personality factors. *Personality and Individual Differences* 45, 2 (2008), 191 – 196. https://doi.org/10.1016/j.paid.2008.03.007
[20] Ryan Rhodes and Nicole E.I. Smith. 2007. Personality Correlates of Physical Activity: A Review and Meta-Analysis. 39 (05 2007), S341.
[21] Jacopo Staiano, Bruno Lepri, Nadav Aharony, Fabio Pianesi, Nicu Sebe, and Alex Pentland. 2012. Friends Don'T Lie: Inferring Personality Traits from Social Network Structure. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*. ACM, New York, NY, USA, 321–330. https://doi.org/10.1145/2370216.2370266
[22] Angelina R Sutin, Luigi Ferrucci, Alan B. Zonderman, and A. Terracciano. 2011. Personality and obesity across the adult life span. *Journal of personality and social psychology* 101 3 (2011), 579–92.
[23] Rui Wang, Weichen Wang, Min S. H. Aung, Dror Ben-Zeev, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A. Scherer, and Megan Walsh. 2017. Predicting Symptom Trajectories of Schizophrenia Using Mobile Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 110 (Sept. 2017), 24 pages. https://doi.org/10.1145/3130976
[24] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. 2018. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 43 (March 2018), 26 pages. https://doi.org/10.1145/3191775
[25] Yanna Weisberg, Colin DeYoung, and Jacob Hirsh. 2011. Gender Differences in Personality across the Ten Aspects of the Big Five. *Frontiers in Psychology* 2 (2011), 178. https://doi.org/10.3389/fpsyg.2011.00178

Appendix D

Conference Poster: MobiSys 2018

The following are the accepted abstract and poster presented at the ACM International Conference on Mobile Systems (MobiSys) 2018. This work includes some of the early studies that lead to the PervasiveHealth submission in Appendix C.

Poster: Inference of Big-Five Personality Using Large-scale Networked Mobile and Appliance Data

Catherine Tong[†], Gabriella M. Harari[§], Angela Chieh[‡],
Otmane Bellahsen[‡], Matthieu Vegreville[‡], Eva Roitmann[‡], Nicholas D. Lane[†*]
[†]University of Oxford, [§]Stanford University, [‡]Nokia Digital Health - Withings, [*]Nokia Bell Labs

ABSTRACT

We present the first large-scale (9270-user) study of data from both mobile and networked appliances for Big-Five personality inference. We correlate aggregated behavioral and physical health features with personalities, and perform binary classification using SVM and Decision Tree. We find that it is possible to infer each Big-Five personality at accuracies of 75% from this dataset despite its size and complexity (mix of mobile and appliance) as prior methods offer similar accuracy levels. We would like to achieve better accuracies and this study is a first step towards seeing how to model such data.

CCS CONCEPTS

• **Human-centered computing** → *Ubiquitous and mobile devices*; *Ubiquitous and mobile devices*; • **Applied computing** → *Psychology*; *Psychology*;

1 INTRODUCTION

Personality traits describe a person's characteristic patterns of thinking, feeling, and behaving. The most widely used model for measuring personality focuses on the Big-Five personality traits - extraversion, agreeableness, conscientiousness, emotionalstability and openness. From an applied perspective, knowing a user's personality could be useful for device customization (e.g. personalization based on psychological characteristics) and for understanding contributing factors to their wellbeing. Traditionally, personality traits are measured using self-report surveys, which can be difficult to scale up in commercial settings. However, mobile sensing technologies permit unobtrusive collection of real-world behavioral patterns [3]. Such technologies may be used to classify personality traits passively, without requiring surveys. Prior research examing this topic [1] indicates the capabilities of using mobile data for their classification. However, these studies used small datasets (under 100 users) and focused on social data (e.g. SMS logs) from only mobile devices. In this study, we aim to understand if methods used in prior studies [1] can be applied to an increasingly popular form of data - large-scale datasets with sparsely collected passive behavioural data.

MobiSys '18, June 10–15, 2018, Munich, Germany

ACM ISBN 978-1-4503-5720-3/18/06.
https://doi.org/10.1145/3210240.3210823

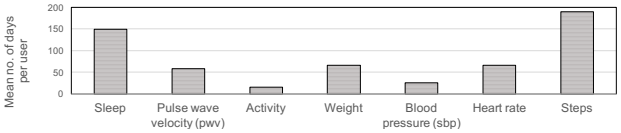


Figure 1: Data per modality. Left to right: modalities the most strongly correlated with Big-Five traits to the least.

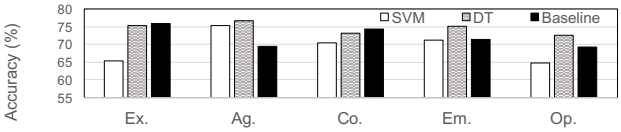


Figure 2: Classification accuracies.

2 METHODOLOGY

Data. 9270 Withings-device users completed a TIPI survey [2] to measure their Big-Five personality traits. We collect sensor data from all Withings devices each user owns (e.g. weight scale, sleep tracker) in daily resolutions over a window of up to 1 year. The data is *passive* with some inferred behavioral statistics (e.g. sleep-stage durations). Each modality is in varying amounts depending on individual usage (Fig. 1).

Data Analysis. We split sensor data across months, aggregating events to extract features, which we use to perform Pearson's correlation with personality traits.

Classification. We consider a binary classification task for each Big-Five personality, using the median to group each into 2 classes. We use features most strongly correlated with personality and at least 500 user-months as input features to SVM and Decision Tree.

3 RESULTS

Correlations. We find a number of strong and significant correlations between features and traits ($|r|>0.1$ and $p<0.01$), e.g. mean and variance in bed-in time with co. and em.; mean weight with ag.

Classification. We use 17 input features from sleep, pwv, weight and sbp. The current study achieves similar accuracies to those observed in [1] (Fig. 2).

Summary. Our study replicates the personality classification analysis done in [1]. We show the application of SVM and Decision Tree onto a new type of large-scale complex dataset to binarily classify Big-Five personality traits is able to achieve accuracies of 74.6%.

REFERENCES

[1] Chittaranjan, G., et al. Who's who with big-five: Analyzing and classifying personality traits with smartphones. *ISWC '11*, 29–36.
[2] Gosling, S., et al. A very brief measure of the big-five personality domains. 504–528.
[3] Harari, G., et al. Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. 838–854.

Inference of Big-Five Personality Using Large-scale Networked Mobile and Appliance Data

Catherine Tong[†], Gabriella M. Harari[§], Angela Chieh[‡], Otmane Bellahsen[‡], Matthieu Vegreville[‡], Eva Roitmann[‡], Nicholas D. Lane[†*]

[†]University of Oxford, [§]Stanford University, [‡]Nokia Digital Health - Withings, *Nokia Bell Labs

UNIVERSITY OF OXFORD
DEPARTMENT OF COMPUTER SCIENCE
NOKIA Bell Labs
EPSRC
Engineering and Physical Sciences Research Council

Introduction

Motivation: Personality inference is useful for device customisation and wellbeing studies.

Big Five Personality: *Extraversion, Agreeableness, Conscientiousness, Emotionalstability*. Measuring through self-report surveys is cumbersome.

Existing Work: Device-based approach for automated personality inference tend to:

- Use small-scaled and homogeneous sample
- Collect only from one mobile source
- Focus on social and mobile behavior.

Goal: Apply methods used in prior studies to a new style of dataset for personality inference.

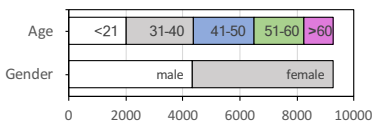
Important Results

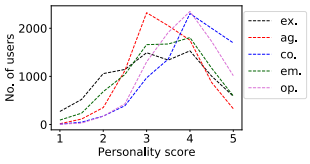
- We replicate the personality classification analysis done in existing work.
- We show the application of SVM and Decision Tree onto a new type of large-scale complex dataset to binarily classify Big-Five

Data

Personality Data:

- TIPI Survey sent out to Withings device users, outputting scores for each personality trait for each respondent
- 9110 valid respondents

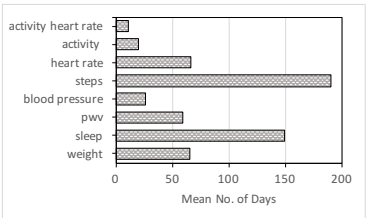




Behavioral and Physiological Data:

Unique Characteristics:

- Large-scale, diverse population
- Mix of data sources
- Mix of raw and inference data
- Sparsity
- Focus on behavioral and physiological features



Framework

Feature Extraction
Aggregating sensor data across months, taking mean and variance for each modality.

Data Analysis
Pearson's Correlation test between features and personality traits.

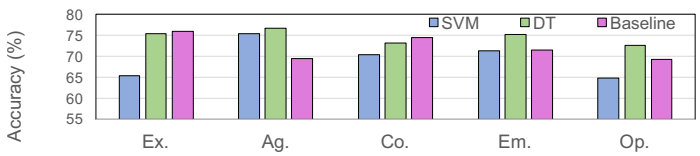
Feature Selection
Found 17 features most strongly correlated with personalities and with at least 500 user-months.

Personality Inference
Binary Classification (grouping each trait by median) using SVM and Decision Tree.
We use Leave-One-Out-Cross-Validation.

Results

Correlation:

We find a number of strong and significant correlations between features and traits ($|r|>0.1$ and $p<0.01$), e.g. mean and variance in bed-in time with co. and em.; mean weight with ag.



Inference:

Using input features from sleep, pwv, weight and sbp, we achieve similar accuracies to baselines observed in *Chittaranjan et al. 2011.*

Future Work

We would like to achieve better accuracies and this study is a first step towards seeing how to model such data.