DISS. ETH NO. 26226

LEARNING TO TREAT, EXPLAIN AND DIAGNOSE WITH NEURAL NETWORKS

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

PATRICK SCHWAB

Dipl.-Ing., University of Vienna

born on 23.01.1991

citizen of Austria

accepted on the recommendation of

Prof. Dr. Walter Karlen, ETH Zurich, examiner
Prof. Dr. Joachim M. Buhmann, ETH Zurich, co-examiner

2019

This page was intentionally left blank.

Abstract

The widespread adoption of wearables, smart devices, and electronic health records systems has led to a surge in available data on personal health and well-being. These novel data sources cover a large amount of factors that influence one's health, offer an objective view of patients at unprecedented temporal resolution, and have the potential to improve healthcare by enabling better decision-making in diagnosing, monitoring, and treating diseases. Machine-learning methods, such as deep learning, have emerged as the state-of-the-art in learning from large-scale health data, but their use for medical tasks is often challenging in practice. Some of the biggest challenges in using deep learning in medicine are that obtaining expert labels is laborious and expensive, that integrating information from multiple data sources over long periods of time is difficult, that interpreting whether black-box models have learned meaningful patterns is in many cases not possible, and that a causal understanding of outcomes is necessary to make treatment recommendations. To address these challenges, we develop several new methods for leveraging distant supervision to reduce the number of required expert annotations, for integrating information from multiple data streams over long periods of time, for jointly producing accurate feature importance scores and predictions, and for estimating individual treatment effects from observational data. We apply our proposed methods to challenging real-world tasks from the medical domain, such as reducing false alarms in critical care, diagnosing Parkinson's disease from smartphone data, identifying discriminatory genes across cancer types, and estimating optimal treatment and dosage choices for mechanical ventilation. Our contributions advance the state-of-the-art in using deep learning for medical applications, and highlight the potential of using machine learning and large-scale health data to improve healthcare.

This page was intentionally left blank.

Zusammenfassung

Der weitverbreitete Einsatz von Wearables, Smart Devices, und elektronischen Gesundheitsakten hat zu einem rasanten Anstieg der verfügbaren Daten im Bereich der persönlichen Gesundheit und des Wohlbefindens geführt. Diese neuartigen Datenquellen decken einen grossen Teil der Faktoren ab, die die persönliche Gesundheit beeinflussen, bieten eine objektive Sicht auf die Gesundheit von Patienten in beispielloser zeitlicher Auflösung, und haben enormes Potential die Gesundheitsversorgung, über bessere Entscheidungen in Diagnose, Krankheitsüberwachung und -behandlung, zu verbessern. Methoden des maschinellen Lernens, wie zum Beispiel Deep Learning, sind der Stand der Technik für das Lernen auf umfangreichen Gesundheitsdatenbanken. Der Einsatz von Methoden des maschinellen Lernens auf Gesundheitsdaten ist jedoch in der Praxis oft herausfordernd. Einige der grössten Herausforderungen im Einsatz von Deep Learning auf Gesundheitsdaten sind, dass Annotationen von medizinischen Experten nur mit viel Zeitaufwand und Kosten gesammelt werden können, dass die Integration von Informationen aus mehreren unterschiedlichen Quellen über lange Zeiträume schwierig ist, dass oft nicht ersichtlich ist, auf welcher Basis die gelernten Modelle ihre Entscheidungen fällen, und, dass Modelle die kausalen Zusammenhänge zwischen den Eigenschaften von Individuen und Behandlungen lernen müssen, um akkurate Behandlungsempfehlungen geben zu können. In dieser Arbeit adressieren wir die genannten Herausforderungen, indem wir neue methodologische Ansätze für neuronale Netze entwickeln, die es ermöglichen, mittels entfernter Überwachung die Anzahl der benötigten Expertenannotationen zu reduzieren, Informationen aus mehreren Quellen über lange Zeitperioden zu integrieren, Modelle zu trainieren, die sowohl Informationen über

die Wichtigkeit der Eingangsdaten als auch akkurate Vorhersagen liefern, und die aufgrund von Beobachtungsdaten mögliche individuelle Behandlungsresultate schätzen können. Wir evaluieren unsere Ansätze auf einer Reihe von herausfordernden Aufgabenstellungen aus dem medizinischen Bereich, wie zum Beispiel der Reduktion von falschen Alarmen auf Intensivstationen, der Diagnose der Parkinson-Krankheit mittels Smartphone-Daten, der Identifikation von Genen, die zwischen verschiedenen Arten von Krebs unterscheiden, und der Schätzung von optimalen Behandlungsschritten und -dosierungen in der mechanischen Beatmung auf der Intensivstation. Diese Arbeit treibt den Stand der Technik im Einsatz von Deep Learning für medizinischen Einsatzzwecke voran und hebt, mit den präsentierten Anwendungen, das Potential zur Verbesserung der Gesundheitsversorgung mittels maschinellem Lernen auf umfassenden Gesundheitsdaten hervor.

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor, Prof. Walter Karlen, whose exceptional advice, guidance, and support has been invaluable throughout my doctoral studies, and to my second examiner, Prof. Joachim M. Buhmann, whose wisdom and support has helped me tremendously in the final phase of my doctoral studies. I am indebted to my colleagues at ETH Zurich, Anita Meinke, Caroline Lustenberger, Djordje Miladinovic, Gaetano Scebba, Jelena Dragas, Jia Zhang, Kanika Dheman, Ku-young Chung, Laura Ferster, Laura Tüshaus, Monica Moreo, and Stefan Bauer. Their presence was the reason that my time at ETH Zurich was always enjoyable. My clinical and industry collaborators, Prof. Emanuela Keller, David J. Mack, Christian Strässle, Carl Muroi, Ronald Stam, Walter O. Frey, Jörg Spörri, Jaap Swanenburg, Lars Lünenburger, and the numerous patients that generously made their health data available for our research deserve a special mention. Without them none of the research presented in this dissertation would have been possible. In addition, I am lucky to have had the opportunity to work with many talented students: Avik Mukhija, Benedikt Dietz, Georgia Channing, Lorenz Linhardt, Matas Pocevicius, and Selin Olenik. Watching them succeed has been an incredibly fulfilling experience. On a more personal note, I would also like to thank my friends back in Vienna for making my frequent visits home memorable. Finally, this work would not have been possible without the unwavering support from my family: Andrea, Heinz, Isabella, Raffaela, Tobias, Erika, and Hans Schwab.

Funding, Infrastructure, and Data. The results presented in this work were partially funded by the Swiss National Science Foundation (SNSF) project No. 150640, and Nos. 167195 and 167302 within the National Research Program (NRP) 75 "Big Data", and the Swiss Commission for Technology and Innovation (CTI) project No. 25531. We acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp graphics processing units (GPUs) used for this research. The data used in our study on diagnosing Parkinson's disease with smartphones were contributed by users of the Parkinson mPower mobile application as part of the mPower study developed by Sage Bionetworks and described in Synapse (doi:10.7303/syn4993293). Contains public sector information licensed under the Open Government Licence (OGL) v3.0. The results shown here are in whole or part based upon data generated by the The Cancer Genomic Atlas (TCGA) Research Network: `http://cancergenome.nih.gov/`. We acknowledge the University of California, Irvine (UCI) Machine Learning Repository (Dua & Graff, 2017) that hosts the New York Times corpus we used in this work. We thank the High-Performance Computing (HPC) group of the Scientific Information Technology Services section at ETH Zurich for providing the compute infrastructure used for several of our experiments.

Contents

This page was intentionally left blank.

List of Figures

List of Tables

List of Abbreviations

AAAI	Association for the Advancement of Artificial Intelligence
ACM	Association for Computing Machinery
ADNI	Alzheimer's Disease Neuroimaging Initiative
ALS	Amyotrophic Lateral Sclerosis
AME	Attentive Mixture of Experts
ARDS	Acute Respiratory Distress Syndrome
ARIMA	Autoregressive Integrated Moving Average
ART	Arterial Blood Pressure
AUC	Area Under the Receiver Operating Characteristic Curve
AUPR	Area Under the Precision Recall Curve
BART	Bayesian Additive Regression Trees
BLSTM	Bidirectional Long Short-term Memory
BN	Batch Normalisation
BNF	British National Formulary
BNN	Balancing Neural Network
BRCA	Breast Carcinoma
CATE	Conditional Average Treatment Effect
CEVAE	Causal Effect Variational Autoencoder
CF	Causal Forests
CFRNET	Counterfactual Regression Networks
CI	Confidence Interval
CMGP	Causal Multi-task Gaussian Processes

CNN	Convolutional Neural Network
COAD	Colon Adenocarcinoma
CPU	Central Processing Unit
CTI	Swiss Commission for Technology and Innovation
DeepLIFT	Deep Learning Important FeaTures
DPE	Mean Dosage Policy Error
DRNet	Dose Response Network
DSMT-Net	Distantly Supervised Multitask Network
EAM	Evidence Aggregation Model
ECG	Electrocardiography
EHR	Electronic Health Records
ETH	Swiss Federal Institute of Technology
FDA	United States Food and Drug Administration
FNN	Feedforward Neural Network
GAN	Generative Adversarial Network
GANITE	Generative Adversarial Nets for inference of Individualised Treatment Effects
GBP	Pound Sterling
GP	General Practice
GPS	Generalised Propensity Score
GPU	Graphics Processing Unit
GVA	Gross Value Added
HPC	High-Performance Computing
ICP	Intracranial Pressure
ICU	Intensive Care Unit
IEEE	Institute of Electrical and Electronics Engineers
ITE	Individual Treatment Effect
KIRC	Kidney Renal Clear Cell Carcinoma

kNN	k-Nearest Neighbours
LIME	Local Interpretable Model-agnostic Explanations
LSTM	Long Short-term Memory
LUAD	Lung Adenocarcinoma
MeSH	Medical Subject Headings
MFCC	Mel-frequency Cepstral Coefficients
MGE	Mean Granger-causal Error
MIMIC	Medical Information Mart for Intensive Care
MISE	Mean Integrated Square Error
MLP	Multi-layer Perception
MNIST	Modified National Institute of Standards and Technology database
MSE	Mean Squared Error
MVICU	Mechanical Ventilation in the Intensive Care Unit
NHS	British National Health Service
NN-MISE	Nearest Neighbour Mean Integrated Square Error
NRP	Swiss National Research Program
OECD	Organisation for Economic Co-operation and Development
OGL	Open Government Licence
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PD	Parkinson's Disease (Chapter 3)
PD	Propensity Dropout (Chapter 5)
PE	Mean Policy Error
PM	Perfect Match
PPG	Photoplethysmogram
PRAD	Prostate Adenocarcinoma
PRO-ACT	Pooled Resource Open-Access ALS Clinical Trials

PSM	Propensity Score Matching
RCT	Randomised Controlled Trial
ReLU	Rectified Linear Unit
RF	Random Forest
RNN	Recurrent Neural Network
SHAP	SHapley Additive exPlanations
SLSQP	Sequential Least Squares Programming
SMAPE	Symmetric Mean Absolute Percentage Error
SNSF	Swiss National Science Foundation
TARNET	Treatment Agnostic Representation Network
TCGA	The Cancer Genomic Atlas
UCI	University of California, Irvine
UK	United Kingdom
US	United States
VAE	Variational Autoencoder

Chapter 1

Introduction

"Curiosity always precedes a problem looking to be solved." - Galileo Galilei

Wearables, smart devices, electronic health records (EHR) systems, healthcare providers, and payers systematically collect large quantities of data about our health and well-being at unprecedented temporal resolution (Jensen et al., 2012; Murdoch & Detsky, 2013). These datasets cover a wide range of factors that influence personal health: Motion sensors on smart devices and wearables, such as accelerometers, gyroscopes and the Global Positioning System, can potentially track a person's physical activities throughout the day (Lara & Labrador, 2013), EHRs and health insurance claims databases store standardised information on treatments, diagnoses, clinical test results and healthcare system utilisation (Kimura et al., 2010; Adler-Milstein et al., 2015), medical imaging provides an in-depth view into body structure and physiology (Suetens, 2017), and genetic testing enables us to screen for biological predispositions (Robson et al., 2015). When combined, these and other data sources may be used to form a comprehensive *digital health profile* that covers many of the factors that influence personal health (Figure 1.1). A natural question to ask in this context is whether and how we can utilise this wealth of information to diagnose diseases, predict disease progression, recommend treatments that lead to better outcomes, and, ultimately, improve decision-making in healthcare.

Contemporary longitudinal health databases frequently contain data on thousands of people, each associated with hundreds to thousands of measured data points. Notable examples of such large longitudinal health

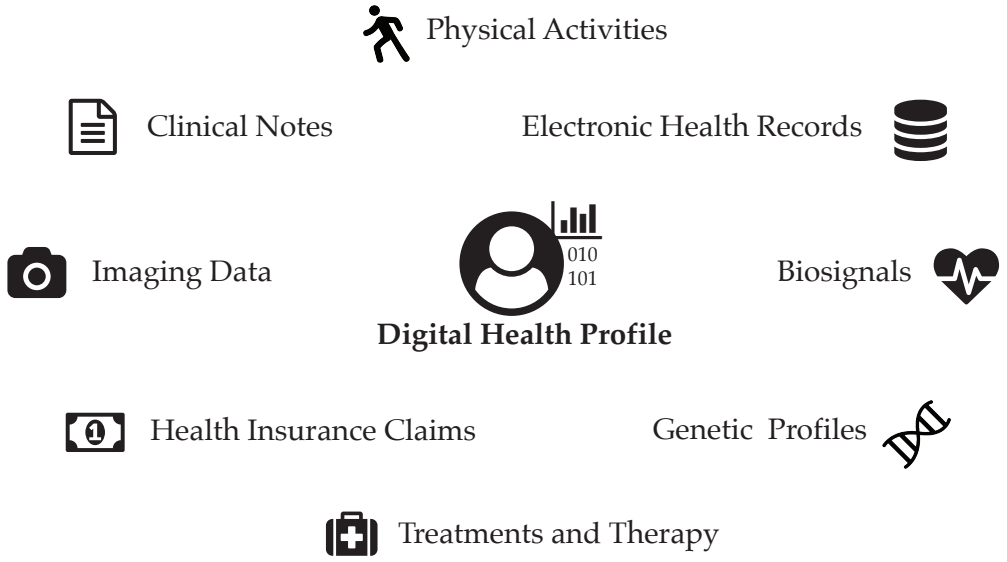


Figure 1.1: Some of the most informative data streams that form an individual's comprehensive digital health profile (Schwab, 2017).

databases are, e.g. the UK Biobank with over half a million participants in the United Kingdom (UK) (Sudlow et al., 2015), the China Kadoorie Biobank that follows half a million adults in China (Chen et al., 2011), the Million Veteran Program with almost 400 thousand enrolled veterans in the United States (US) (Gaziano et al., 2016), and the Medical Information Mart for Intensive Care (MIMIC) database that covers 50 thousand patients that were admitted to intensive care units in the US (Saeed et al., 2011; Johnson et al., 2016). In addition to these country-scale databases, there also exist dozens of relatively smaller, more specialised databases that focus on people with a particular disorder. Examples of such specialised databases are the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Petersen et al., 2010), the mPower study for smartphone monitoring in Parkinson's disease (Bot et al., 2016), the SchizConnect database mediator for schizophrenia and related disorders (Wang et al., 2016), the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) platform with over 1 800 amyotrophic lateral sclerosis (ALS) patients (Küffner et al., 2015), and The Cancer Genomic Atlas (TCGA) project that collected clinical and gene expression data from several types of cancers in 9 659 individuals (Weinstein et al., 2013). These repositories contain vast amounts of information that could potentially

be used to better understand how medical treatments, behaviors, mental, social, environmental, and hereditary factors influence personal health and well-being (Krumholz, 2014).

However, it is not feasible for medical experts to interpret datasets of this scale manually. Computational methods that scale to large numbers of input features and data points have therefore emerged as the de-facto state-of-the-art for learning from large health databases. Machine learning, the study of inferring and recognising patterns from data, is concerned with developing, researching, and evaluating such computational methods. Driven by the triumvirate of readily available large-scale datasets, scalable software implementations (Chen et al., 2015; Chollet et al., 2015; Abadi et al., 2016; Paszke et al., 2017) and efficient compute hardware (LeCun et al., 2015; Jordan & Mitchell, 2015), the field of machine learning has recently experienced what could arguably be described as a golden age, with several breakthrough results achieved in important fields such as object recognition (Deng et al., 2009; Szegedy et al., 2017), game-playing agents (Mnih et al., 2015; Silver et al., 2016), and generative modelling (Goodfellow et al., 2014; Karras et al., 2017). In healthcare, similarly notable results have been achieved in the field of medical image processing, where researchers used machine learning to, for example, reach human-level performance in skin cancer classification (Esteva et al., 2017), detection of mammographic lesions (Kooi et al., 2017), diagnosing based on chest radiographs (Rajpurkar et al., 2018), and in recommending referrals for retinal diseases based on optical coherence tomography (De Fauw et al., 2018).

The machine-learning approach behind the above-mentioned breakthrough results is *deep learning*, in particular with deep neural networks. Deep learning is a representation-learning method that, in essence, utilises multiple, increasingly complex layers of abstraction to learn arbitrarily intricate functions (Schmidhuber, 2015; LeCun et al., 2015; Goodfellow et al., 2016). For example, when training a model to recognise objects in images, lower levels of abstraction, i.e. those closer to the input data, would correspond to object edges, whereas higher layers would correspond to parts of objects, such as the wheels of a car (Goodfellow et al., 2016). Due to their

remarkable ability to learn to extract meaningful features from raw feature representations without manual feature engineering, deep neural networks have become the go-to method for learning from raw and unstructured data, such as images, texts, and time series. These properties make neural networks well-suited for applications in medicine as well, as performing medical tasks often involves data that is either of temporal, for example sensor data from wearables (Clifton et al., 2012, 2013), biosignal monitors (Schwab et al., 2018a) or EHRs, textual, such as clinical notes (Grnarova et al., 2016), or spatial nature, such as medical imaging data (Ching et al., 2018).

In addition to its ability to extract rich feature representations from raw data, deep learning is also a highly flexible framework that may be extended to account for the domain-specific challenges in using machine learning for medical applications (Miotto et al., 2017; Ching et al., 2018). Some of the main challenges in applying deep learning to medical tasks are that obtaining ground-truth labels from medical experts is laborious and expensive (Ching et al., 2018), that integrating information from multiple data sources over long periods of time is difficult (Miotto et al., 2017), that interpreting the decisions of black-box models is in many cases not possible (Miotto et al., 2017; Ching et al., 2018), and that causal inference of counterfactual outcomes is necessary to make treatment recommendations from observational data (Ghassemi et al., 2018; Schwab et al., 2019a).

In this work, we address these challenges by developing novel methodological approaches to using deep learning on large-scale health data. In addition, we demonstrate the potential of these methods by applying them to a number of challenging tasks in the medical domain (Table 1.1). In particular, we develop new approaches for leveraging distant supervision from multiple auxiliary tasks to reduce the number of required expert annotations (Chapter 2), for learning hierarchical models that both assess individual data streams as well as integrate data from multiple heterogenous data streams over time (Chapter 3), for learning to jointly produce accurate estimates of feature importance and predictions in a single neural network (Chapter 4), and for learning to estimate individual dose-response from observational data when multiple treatment options with associated dosage parameters

are available (Chapter 5). We evaluate our proposed methods on several tasks from the medical domain, such as reducing false alarms in critical care, diagnosing Parkinson's disease from smartphone data, identifying discriminatory genes across cancer types, and estimating optimal treatment choices for mechanical ventilation in critical care and cancer treatment recommendation. This work advances the state-of-the-art in using deep learning for medical applications, and highlights the potential of using machine learning and large-scale health data to improve healthcare.

1.1 Research Goals

Our primary research objective is to advance the state-of-the-art in the use of deep neural networks and large-scale health databases for medical applications. Specifically, our goals are to make methodological contributions that enable the use of deep learning on health databases to perform diagnoses, to explain model decisions, and to recommend optimal treatment choices and dosages. We additionally aim to evaluate these methodological contributions on real-world tasks from the medical domain. Formally, our overarching research question can be stated as:

How can we utilise deep learning and observational health data to provide diagnoses, explanations, and treatment recommendations for medical applications?

We address this question in four parts, where each part addresses one key challenge in the application of deep neural networks in medicine.

(Q1) How can we improve the predictive performance of neural networks when labels are scarce but large amounts of unlabelled time series data are available?

Our research hypothesis for (Q1) is that we can utilise distant supervision from multiple automatically inferred auxiliary tasks to improve predictive performance when large amounts of unlabelled time series data are available. In addition, we hypothesise that task-specific auxiliary tasks are better suited for distantly supervised multitask learning than task-unspecific auxiliary tasks, such as the reconstruction-based objectives used

in Variational Auto-Encoders (VAEs) (Kingma et al., 2014) and Ladder Networks (Rasmus et al., 2015). Answering this question is particularly important for healthcare applications, because the time of medical experts is limited and expensive, and large-scale annotation efforts are therefore in many cases not feasible. Large amounts of unlabelled, temporal data are, however, often readily available, e.g. from continuous monitoring.

(Q2) How can we utilise deep learning to better integrate information from multiple heterogenous, raw data streams over long periods of time?

For (Q2), we hypothesise that a hierarchical approach that separates learning to assess each respective individual data stream, and learning to integrate data from these high-level assessments over time into two distinct stages may be used to better integrate information from multiple heterogenous, raw data sources over long periods of time. The reasoning behind this hypothesis is that the learning problem is likely easier when considering each of these tasks on their own rather than attempting to learn both tasks end-to-end. This research question is particularly relevant for the medical domain, because the comprehensive monitoring of patients' symptoms frequently involves collecting several types of heterogenous measurements over long periods of time.

(Q3) How can we train deep neural networks to jointly produce predictions and accurate estimates of feature importance?

Our hypothesis for (Q3) is that techniques from causal inference can be used to quantify to what degree a specific input causes a given output of a deep neural network. We reason that such a causal learning objective could be used to train a neural network to jointly produce both accurate estimates of feature importance for individual samples, and predictions for the main task. We hypothesise that this approach would likely be significantly more efficient in terms of computation time than existing model interpretation techniques based on sensitivity analysis, because model interpretation techniques based on sensitivity analysis scale poorly to large numbers of input features, whereas a neural network trained to jointly produce accurate estimates of feature importance and predictions would provide estimates

of feature importance essentially for free along with the predictions. The availability of feature importance estimates for individual predictions is especially important for applications in high-stakes environments, such as medicine, because it enables the understanding, interpretation and validation of model outputs (Doshi-Velez & Kim, 2017).

(Q4) How can we train deep neural networks to learn to estimate individual treatment effects from observational data when there are multiple treatments with associated continuous dosage parameters available?

Based on prior research on the importance of model structure for learning to estimate individual treatment effects (Shalit et al., 2017; Schwab et al., 2018b; Alaa & Schaar, 2018) and under unconfoundedness assumptions (Imbens, 2000; Lechner, 2001), we hypothesise that a hierarchical model architecture specifically tailored for estimating treatment effects for multiple treatments with associated continuous dosage parameters could enable us to better learn to estimate individual treatment effects in the setting outlined by (Q4). Determining the right treatment option and dosage for individual patients is one of the most important tasks in the medical domain, and physicians typically rely on evidence from randomised controlled trials to decide which treatment options are most likely to be efficacious for a given patient. However, conducting randomised controlled trials is expensive, time-consuming, in many cases not feasible for ethical reasons (Schafer, 1982), and only yields information about the average treatment effect across a whole population rather than for individuals (Schwab et al., 2019a). Better methods to learn to estimate individual treatment effects from observational data are therefore necessary to achieve the key promise of precision medicine (Collins & Varmus, 2015) - to deliver the right treatment to the right patient at the right time (Miotto et al., 2017).

Scope. The focus of this work lies primarily in addressing the technical challenges associated with the use of deep learning in the medical domain. Besides the technical challenges, there exist a number of non-technical challenges in learning from healthcare data. Examples of open non-technical challenges of machine learning in medicine include, among others,

its societal implications (Dzau & Balatbat, 2018), questions of ethics and fairness (Darcy et al., 2016; Char et al., 2018; Beam & Kohane, 2018; Vayena et al., 2018), data sharing and privacy (Murdoch & Detsky, 2013; Horvitz & Mulligan, 2015; Mooney & Pejaver, 2018; Shameer et al., 2018), how predictive algorithms in medicine should be regulated (Yaeger et al., 2019; Parikh et al., 2019), how the adoption of machine learning in the healthcare system will impact clinical practice (Obermeyer & Emanuel, 2016), and how adversarial attacks on medical machine learning systems could impact the healthcare system (Finlayson et al., 2019). These questions are in many cases as or even more important than the technical challenges, and we therefore remain cognisant of them throughout this thesis. However, this work does not explicitly address the non-technical challenges associated with the use of machine learning in medicine.

1.2 Contributions

This dissertation is an interdisciplinary work and, as such, contains scientific contributions both in terms of methodological advances in the field of machine learning as well as in the application of machine-learning methods to challenging tasks from the medical domain.

The main methodological contributions of this thesis are as follows:

Chapter 2 We introduce a novel neural network architecture for learning from multiple distant auxiliary tasks, and present a method for automatically inferring a large number of auxiliary tasks for distantly supervised multitask learning.

Chapter 3 We develop a hierarchical machine-learning approach to learning to integrate information from multiple heterogenous, raw data streams covering long periods of time, and introduce a hierarchical neural attention mechanism that quantifies the importance of both individual data streams and segments within those data streams.

Table 1.1: Overview of the challenges in deep learning in medicine addressed in this dissertation, the proposed approaches, and the presented applications of these approaches.

	Challenge	Proposed Approach	Presented Applications
Chapter 2	Few expert labels	Distantly supervised multitask learning	Alarm reduction in critical care
Chapter 3	Temporal modelling	Evidence Aggregation Model	Diagnosing Parkinson's disease from smartphone data
Chapter 4	Interpretability of neural networks	Jointly learning to predict, and estimate feature importance	Identifying discriminatory genes across cancer types, determining factors driving medical prescription demand
Chapter 5	Counterfactual inference with multiple treatments and continuous dosages	Dose Response Networks	Mechanical ventilation in critical care, recommendations for cancer treatment

Chapter 4 We describe a mixture of experts architecture that uses attentive gating to assign weights to individual experts, and present a secondary Granger-causal objective that can be used to train neural networks to jointly learn to produce accurate estimates of feature importance and predictions.

Chapter 5 We introduce a method for learning to estimate counterfactual outcomes for multiple available treatment options with associated continuous dosage parameters with neural networks, and develop performance metrics, model selection criteria, and open benchmarks for estimating individual dose-response curves.

In terms of medical applications, our main contributions are:

Chapter 2 We develop and evaluate a deep-learning approach for reducing the number of false alarms in critical care.

Chapter 3 We develop and evaluate a deep-learning approach for learning to diagnose Parkinson's disease from smartphone data.

Chapter 4 We develop and evaluate a deep-learning approach for learning to determine drivers of medical prescription demand, and to identify discriminatory genes across several types of cancer.

Chapter 5 We develop and evaluate, in semi-synthetic experiments, a deep-learning approach for learning to estimate optimal treatment and dosage choices for mechanical ventilation in critical care, and cancer treatment recommendation.

1.3 Thesis Outline

In the following paragraphs, we outline the structure of this thesis in detail, and describe its content both in terms of methodological contributions as well as the presented applications.

In Chapter 2, we introduce a novel approach to distantly supervised multitask learning from multiple high-resolution biosignal monitoring systems based on neural networks. We apply this approach to false alarm reduction in critical care, and demonstrate that our approach compares favourably to several state-of-the-art methods for semi-supervised learning on a large-scale dataset collected by our clinical collaborators at the Neurocritical Care Unit of the University of Zurich, Switzerland that contains almost 14 000 clinical alarms. With hundreds of alarms triggered in intensive care units per patient per day and a reported false alarm rate of up to almost 90% (Drew et al., 2014), alarm desensitisation is an important clinical problem that could in the future potentially in part be addressed through the use of intelligent algorithms for false alarm reduction.

In Chapter 3, we present a machine-learning approach to integrating data from multiple heterogenous smartphone-based tests, including walking, voice, memory and tapping tests, that are performed regularly over long periods of time. We apply this approach to the task of diagnosing Parkinson's disease using smartphone data from a cohort of more than 1 800 people with and without Parkinson's disease - the largest cohort for evaluating a smartphone-based approach to diagnosing Parkinson's disease to date (Schwab & Karlen, 2019b). Our experimental results show that an end-to-end implementation of our approach with recurrent neural networks and hierarchical neural soft attention achieves a better predictive performance at diagnosing Parkinson's disease from smartphone data than several strong baselines. We also qualitatively analyse the patterns identified by our model by visualising the attention factors assigned by the hierarchical neural attention mechanism both at the level of all tests as well as at the level of individual tests. With around 25% of diagnoses being incorrect (Pahwa & Lyons, 2010; Schwab & Karlen, 2019b), Parkinson's disease is difficult to diagnose even for medical experts. Smartphones-based tests and machine learning have the potential to be used as additional digital biomarkers for the diagnosis of Parkinson's disease in the future.

In Chapter 4, we describe a new approach to jointly training neural networks to produce both accurate estimates of feature importance and

predictions using a labelled dataset. Our approach is based on a secondary Granger-causal objective that may be used to optimise neural networks to produce accurate estimates of feature importance for individual samples. We perform an extensive experimental evaluation on three datasets, including quantifying important features in handwritten digits, determining the drivers of medical prescription demand, and identifying discriminatory genes across several types of cancer. Our results show that our approach is competitive with existing state-of-the-art methods for estimating feature importance while being orders of magnitude faster to compute. Interpretable models are important for many machine-learning tasks, particularly in medicine, where the requirements for understanding, validating, and interpreting a predictive model's outputs are high.

In Chapter 5, we present a method for learning to estimate individual dose-response from observational data using neural networks. Our method combines a specialised model structure with extensions of several existing regularisation strategies for addressing the treatment assignment bias inherent in observational data to the dose-response setting. To evaluate our method, we introduce three performance metrics and three benchmarks for estimating counterfactual outcomes in settings with multiple treatment options with associated dosage parameters. In an extensive experimental evaluation, we show that our method outperforms several existing state-of-the-art methods in estimating counterfactual outcomes from observational data. Accurate machine-learning models for predicting potential counterfactual outcomes under different treatments and dosages could potentially enable us, under certain assumptions, to estimate optimal treatment policies when experimental data is not available. These treatment policies could be used to answer counterfactual questions such as "Which treatment at what dosage should we apply to achieve the optimal outcome?" and, thus, provide actionable treatment recommendations in medical applications.

Lastly, in Chapter 6, we summarise our main findings and discuss their importance in relation to the presented applications and research objectives. We also provide a brief outlook on how the field of machine learning in medicine might develop in the future.

1.4 Publications

This dissertation primarily covers the contents of four published works that are respectively enclosed in Chapters 2, 3, 4 and 5. The publications have previously appeared as (in order of appearance):

(P1) **Schwab, Patrick**, Keller, Emanuela, Muroi, Carl, Mack, David J., Strässle, Christian and Karlen, Walter. Not to cry wolf: Distantly supervised multitask learning in critical care. *International Conference on Machine Learning*, Stockholm, Sweden, 10-15 July, 2018

(P2) **Schwab, Patrick** and Karlen, Walter. PhoneMD: Learning to diagnose Parkinson's disease from smartphone data. *AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, United States, 27 Jan - 1 Feb, 2019

(P3) **Schwab, Patrick**, Miladinovic, Djordje and Karlen, Walter. Granger-causal attentive mixtures of experts: Learning important features with neural networks. *AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, United States, 27 Jan - 1 Feb, 2019

(P4) **Schwab, Patrick**, Linhardt, Lorenz, Bauer, Stefan, Buhmann, Joachim. M., and Karlen, Walter. Learning counterfactual representations for estimating individual dose-response curves. *arXiv preprint arXiv:1902.00981 (in submission)*, 2019

We edited the four enclosed publications in order to standardise the formatting for their inclusion in this thesis. For completeness, we also appended the respective supplementary materials to all of the works enclosed in this dissertation. As is common in research, the works enclosed in this dissertation were produced in a collaborative effort. The contributions of the other named authors of the four enclosed publications were as follows: Emanuela Keller, Carl Muroi, David J. Mack and Christian Strässle devised, implemented and supervised the data collection and alarm annotation at the Neurocritical Care Unit of the University of Zurich, Switzerland and edited publication (P1). For

publication (P4), Lorenz Linhardt contributed source code and evaluations for initial implementations of the News and Mechanical Ventilation in the Intensive Care Unit (MVICU) benchmarks as described in his Master's thesis (Linhardt, 2018), and edited the final manuscript. Djordje Miladinovic edited publication (P3), Stefan Bauer edited and supervised publication (P4), and Joachim M. Buhmann supervised publication (P4). Walter Karlen edited and supervised all enclosed works.

In addition to the four main publications, the following five works were also completed over the course of my doctoral research, and are relevant to this dissertation. However, their contents are not covered in this thesis.

(P5) **Schwab, Patrick**, Scebba, Gaetano C., Zhang, Jia, Delai, Marco and Karlen, Walter. Beat by beat: Classifying cardiac arrhythmias with recurrent neural networks. *Computing in Cardiology*, Rennes, France, Sept 24-27, 2017

(P6) **Schwab, Patrick**, Linhardt, Lorenz and Karlen, Walter. Perfect Match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656 (in submission)*, 2018

(P7) Muroi, Carl, Meier, Sandro, De Luca, Valeria, Mack, David J., Strässle, Christian, **Schwab, Patrick**, Karlen, Walter and Keller, Emanuela. Automated false alarm reduction in a real-life intensive care setting using motion detection. *Neurocritical Care*, 2019

(P8) **Schwab, Patrick**, and Karlen, Walter. A deep learning approach to diagnosing multiple sclerosis from smartphone data. *(in submission)*, 2019

(P9) **Schwab, Patrick**, and Karlen, Walter. CXPlain: Causal explanations for model interpretation under uncertainty. *Advances in Neural Information Processing Systems (to appear)*, 2019

Furthermore, the following two Master's theses were completed within the context of the research presented in this dissertation:

- Pocevicius, Matas (advised by Karlen, Walter and **Schwab, Patrick**). Intelligent decision support for diagnosis and monitoring of Parkinson's disease. Master's thesis, ETH Zurich, Switzerland, 2018
- Linhardt, Lorenz (advised by Buhmann, Joachim M., Karlen, Walter and **Schwab, Patrick**). Learning counterfactual representations for ventilation in critical care: Methods and benchmarks. Master's thesis, ETH Zurich, Switzerland, 2018

This page was intentionally left blank.

Chapter 2

Distantly Supervised Multitask Learning in Critical Care

"Intelligence is not to make no mistakes, but quickly to see how to make them good." - Bertolt Brecht

Patients in the intensive care unit (ICU) require constant and close supervision. To assist clinical staff in this task, hospitals use monitoring systems that trigger audiovisual alarms if their algorithms indicate that a patient's condition may be worsening. However, current monitoring systems are extremely sensitive to movement artefacts and technical errors. As a result, they typically trigger hundreds to thousands of false alarms per patient per day - drowning the important alarms in noise and adding to the exhaustion of clinical staff. In this setting, data is abundantly available, but obtaining trustworthy annotations by experts is laborious and expensive. We frame the problem of false alarm reduction from multivariate time series as a machine-learning task and address it with a novel multitask network architecture that utilises distant supervision through multiple related auxiliary tasks in order to reduce the number of expensive labels required for training. We show that our approach leads to significant improvements over several state-of-the-art baselines on real-world ICU data, and provide new insights on the importance of task selection and architectural choices in distantly supervised multitask learning.

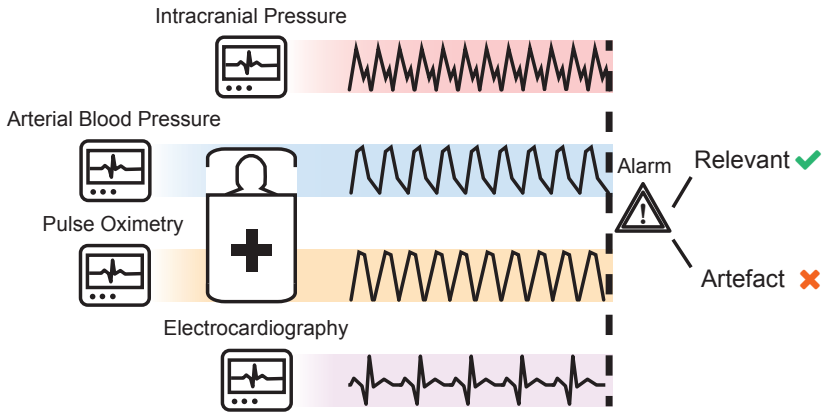


Figure 2.1: Schematic overview of the described problem setting in critical care. Before an alarm is brought to the attention of clinical staff, an alarm classification algorithm could analyse a recent window of the full set of available data streams in order to identify whether the alarm was likely caused by an artefact or technical error and may therefore be reported with a lower degree of urgency.

2.1 Introduction

False alarms are an enormous mental burden for clinical staff and are extremely dangerous to patients, as alarm fatigue and desensitisation may lead to clinically important alarms being missed (Drew et al., 2014). Reportedly, several hundreds of deaths a year are associated with false alarms in patient monitoring in the United States alone (Cvach, 2012)[1].

An intelligent alarm classification system could potentially reduce the burden of a large subset of those false alarms by assessing which alarms were likely caused by either an artefact or a technical error and reporting those alarms with a lower degree of urgency (Figure 2.1). Roadblocks that have so far prevented the adoption of machine learning for this

[1]To the best of our knowledge, there are currently no studies that precisely quantify the effect of alarm fatigue on patient mortality. Quantifying the total number of patient deaths due to alarm fatigue is difficult as they are likely underreported (Sendelbach & Funk, 2013). A search in a database maintained by the United States Food and Drug Administration (FDA) revealed 566 alarm-related death reports from 2005 to 2008 (Honan et al., 2015).

task are the heterogeneity of monitoring systems, the requirement for an extremely high specificity, to avoid suppressing important alarms, and the prohibitively high cost associated with obtaining a representative set of clinically validated labels for each of the manifold alarm types and monitoring system configurations in use at hospitals.

We present a semi-supervised approach to false alarm reduction that automatically identifies and incorporates a large amount of distantly supervised auxiliary tasks in order to significantly reduce the number of expensive labels required for training. We demonstrate, on real-world ICU data, that our approach is able to correctly classify alarms originating from artefacts and technical errors better than several state-of-the-art methods for semi-supervised learning when using just 25, 50 and 100 labelled samples. Besides their importance for clinical practice, our results highlight the power of distant multitask supervision as a flexible and effective tool for learning when unlabelled data are readily available, and shed new light on semi-supervised learning beyond low-resolution image benchmark datasets.

Contributions. In this chapter, we present the following contributions:

- We introduce Distantly Supervised Multitask Networks (DSMT-Nets): A novel neural architecture built on the idea of utilising distant supervision through multiple auxiliary tasks in order to better harness unlabelled data.
- We present a methodology for selecting a large set of related auxiliary tasks in time series, and a training procedure that counteracts adverse gradient interactions between auxiliary tasks and the main task.
- We perform extensive quantitative experiments on a real-world ICU dataset consisting of almost 14 000 alarms in order to evaluate the relative classification performance and label efficiency of DSMT-Nets compared to several state-of-the-art methods.

2.2 Related Work

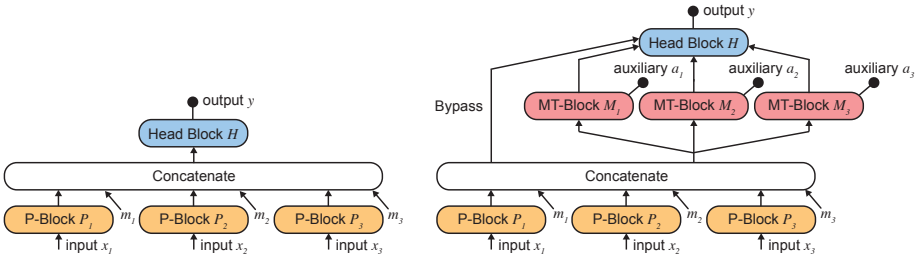
Background. Driven by widespread efforts to automate patient monitoring, there has been a recent surge in works applying machine learning to the vast amounts of data generated in ICUs. One notable driver is the MIMIC (Saeed et al., 2011) dataset that has made ICU data accessible to a large number of researchers. Related works have, for example, explored the use of ICU data for tasks such as mortality modelling (Ghassemi et al., 2014), illness assessment and forecasting (Ghassemi et al., 2015), diagnostic support (Lipton et al., 2016a), patient state prediction (Cheng et al., 2017) and learning weaning policies for mechanical ventilation (Prasad et al., 2017). Applying machine-learning approaches to clinical and physiological data is challenging, because it is heterogenous, noisy, confounded, sparse and of high temporal resolution over long periods of time. These properties are in stark contrast to many of the benchmark datasets that machine-learning approaches are typically developed and evaluated on. Several works therefore deal with adapting existing machine-learning approaches to the idiosyncrasies of clinical and physiological data, such as missingness (Lipton et al., 2016b; Che et al., 2018), long-term temporal dependencies (Choi et al., 2016a), noise (Schwab et al., 2017), heterogeneity (Libbrecht & Noble, 2015) and sparsity (Lasko et al., 2013). We build on several of these innovations in this work.

Alarm Fatigue. The PhysioNet 2015 challenge on false alarm reduction in electrocardiography (ECG) monitoring (Clifford et al., 2015) was one of the most notable efforts to date to address the issue of false alarms in physiological monitoring. Within the challenge, researchers introduced several effective approaches to reducing the false alarm rate of arrhythmia alerts in ECGs (Fallet et al., 2015; Eerikäinen et al., 2015; Krasteva et al., 2016; Plesinger et al., 2016). However, clinicians in the ICU do not just monitor for arrhythmias, but many adverse events at once using a multitude of different monitoring systems. Typically, these monitoring systems operate in isolation on a single biosignal and trigger their own distinct sets of alarms. Previous research has shown that there is an opportunity to use data from

other biosignals to identify false alarms in related waveforms (Aboukhalil et al., 2008). We therefore believe that a comprehensive solution to alarm fatigue requires an approach that accounts for the monitoring setup as a whole, rather than targeting specific systems or alarms in isolation.

Distant Supervision and Multitask Learning. Multitask learning has a rich history in healthcare applications and has, for example, been used for risk prediction in neonatal intensive care (Saria et al., 2010), drug discovery (Ramsundar et al., 2015) and prediction of Clostridium difficile (Wiens et al., 2016). A way of leveraging multitask learning to improve label-efficiency is to learn jointly from complementary unsupervised auxiliary tasks along with the supervised main task. Existing literature refers to the concept of applying indirect supervision through auxiliary tasks, be it for label-efficiency or additional predictive performance, as weak supervision (Papandreou et al., 2015; Oquab et al., 2015), distant supervision (Zeng et al., 2015; Deriu et al., 2017) or self-supervision (Fernando et al., 2017; Doersch & Zisserman, 2017). In particular, (Doersch & Zisserman, 2017) used distantly supervised multitask learning to increase predictive performance in computer vision with up to four hand-engineered auxiliary tasks. Using auxiliary tasks in addition to a main task has also been shown to be a promising approach in reinforcement (Jaderberg et al., 2017; Aytar et al., 2018) and adversarial learning (Salimans et al., 2016). Recently, (Laine & Aila, 2017) proposed to use outputs from the same model at different points in training and with varying amounts of regularisation as additional unsupervised targets for the main task. In contrast to existing works, we present the first approach to distantly supervised multitask learning that automatically identifies a large set of related auxiliary tasks from multivariate time series to jointly learn from labelled and unlabelled data. In addition, our approach scales to hundreds of auxiliary tasks in an end-to-end trained neural network.

Semi-supervised Learning. Beside distant supervision, other state-of-the-art approaches to semi-supervised learning in neural networks include, broadly, methods based on (i) reconstruction objectives, such as Variational Autoencoders (VAEs) (Kingma et al., 2014) and Ladder Networks (Rasmus



(a) DSMT-Net-0 (b) DSMT-Net-3

Figure 2.2: Two DSMT-Net architectures: (a) one without (DSMT-Net-0) and (b) one with three (DSMT-Net-3) auxiliary tasks. The number of horizontally aligned multitask blocks M_j (MT-Blocks; red) is variable. Each multitask block hosts its own auxiliary task a_j. An additional bypass connection gives the head block H (blue) direct access to the concatenated hidden states of the perception blocks P_i (P-Blocks; orange). Each perception block operates on its own input data stream x_i. The model incorporates binary missing indicators m_i for each perception block to handle situations where input data streams are missing.

et al., 2015), and (ii) adversarial learning (Springenberg, 2016; Dai et al., 2017; Li et al., 2017). However, with standard benchmarks consisting primarily of low-resolution image datasets, it is yet unclear to what degree these method's results generalise to heterogenous, long-term and high-resolution time series datasets with informative missingness, as commonly encountered in healthcare applications.

2.3 Distantly Supervised Multitask Networks

Distantly Supervised Multitask Networks (DSMT-Nets) are end-to-end trained neural networks that process n heterogenous input data streams x_i in order to solve a multitask learning problem with one main task and k auxiliary tasks designed to augment the main task. Conceptually, a DSMT-Net consists of the following components: One perception block P_i with $i \in [1 \,..\, n]$ for each of the n input data streams x_i, a variable number k of multitask blocks M_j with $j \in [1 \,..\, k]$, and a single head block H

(Figure 2.2b). Each of these block types is itself a neural network with its own parameters and arbitrary architectures and hyperparameters. The role of the perception blocks P_i is to extract a hidden feature representation $h_{p,i}$ from their respective input data streams x_i:

$$h_{p,i} = P_i(x_i) \tag{2.1}$$

We separate the perception blocks by input stream x_i in order to be able to model a dynamic set of potentially missing input data streams. To allow our model to learn missingness patterns, we follow (Lipton et al., 2016b) and accompany each perception block with a missing indicator m_i that is set to 0 if the data stream x_i is present and 1 if it is missing. We additionally initialise the features $h_{p,i}$ of the missing perception blocks to 0. We then concatenate the features $h_{p,i}$ extracted from the perception blocks and the corresponding missing indicators m_i into a joint feature representation P_c over all input data streams:

$$P_c = \text{concatenate}(h_{p,1}, m_1, ..., h_{p,n}, m_n) \tag{2.2}$$

The joint feature representation P_c combines the information from all feature representations of the input data streams and serves as input to the higher level multitask blocks and the head block. The main role of multitask blocks M_j is to host auxiliary tasks a_j. All multitask blocks are aligned in parallel in order to minimise the distance gradients have to propagate through both to the joint feature representation P_c and from the head block. As output, each multitask block produces a hidden high-level feature representation $h_{m,j}$:

$$h_{m,j} = M_j(P_c) \tag{2.3}$$

Compared to the straightforward approach of directly appending the auxiliary tasks to the head block H, the positioning of multitask blocks below the head block achieves separation of concerns. In DSMT-Nets, the head block focuses on learning a hidden feature representation that is optimised solely for the main task rather than being forced to learn a joint

feature representation that performs well on multiple, possibly competing tasks.

The head block H computes the final model output y and further processes the hidden feature representations $h_{m,j}$ of the multitask blocks via a combinator function (equation (2.4)). In addition to the hidden feature representations of the multitask blocks, the head block retains direct access to P_c via a bypass connection. We motivate the inclusion of a bypass connection with the desire to learn hidden feature representations in multitask blocks that add information over P_c (He et al., 2016). Mathematically, we formulate the head block H as follows:

$$y = H(\text{combine}_{\text{MLP}}(P_c, h_{m,1}, ..., h_{m,k})) \tag{2.4}$$

We note that the DSMT-Net architecture without any multitask blocks corresponds to a naïve supervised neural network over a mixture of expert networks (Jordan & Jacobs, 1994; Shazeer et al., 2017; Schwab et al., 2019b) for each input data stream x_i (DSMT-Net-0; Figure 2.2a).

Combinator Function. In DSMT-Nets, the combinator function integrates $m+1$ data flows from the m multitask blocks' hidden representations as well as the joint feature representation P_c. We propose a combinator function ($\text{combine}_{\text{MLP}}$) that consists of a single hidden-layer multi-layer perceptron (MLP) with a dimensionality twice as big as a single multitask block's feature representation. As input, the MLP receives the concatenation of all the feature representations to be integrated:

$$\text{combine}_{\text{MLP}} = \text{MLP}(\text{concatenate}(P_c, h_1, ..., h_m)) \tag{2.5}$$

2.3.1 Selection of Auxiliary Tasks

One of the most important questions in distantly supervised learning is how to identify suitable auxiliary tasks. A common choice of auxiliary task for un- and semi-supervised learning is reconstruction over the feature and/or hidden representation space. Several modern semi-supervised methods

take this approach (Vincent et al., 2008; Kingma & Welling, 2014; Kingma et al., 2014; Rasmus et al., 2015). Reconstruction is a convenient choice of auxiliary task because it is generically applicable to any input data, neural architecture and predictive task. However, given recent empirical successes by distant supervision with specifically engineered auxiliary tasks (Oquab et al., 2015; Deriu et al., 2017; Doersch & Zisserman, 2017), we reason that (i) more "related" tasks might be a better choice of auxiliary task for semi-supervised learning than reconstruction (Ben-David & Schuller, 2003) and that (ii) using multiple diverse auxiliary tasks might be more effective than just one (Baxter, 2000). Since a predictive feature for a main task is also a good auxiliary task for learning shared predictive representations (Ando & Zhang, 2005), we follow a simple two-step feature selection methodology (Christ et al., 2016) to automatically identify a large set of auxiliary tasks that are closely related to the main task:

1. We extract features from a large pool of manually-designed features from each input time series. Due to the large wealth of research in manual feature engineering, there exist vast repositories of such features for many data modalities, e.g. (Christ et al., 2016). For time series, examples of such features would be, e.g., the autocorrelation at different lag levels or the power spectral density over a specific frequency range.

2. We statistically test the extracted features for their importance related to the main task in order to rank the features by their estimated predictive potential and determine their relevance. A suitable statistical test is, for example, a hypothesis test for correlation between the labels y_{true} and the extracted features using Kendall's τ (Kendall, 1945).

Using this approach, we are able to identify a large, ranked list of predictive features suitable for use as target labels for auxiliary tasks a_j in DSMT-Nets. There are two approaches to choosing a subset of those features as auxiliary targets: (i) in order of feature importance or (ii) randomly out of the set of relevant features. The main difference between the two approaches is that

random selection has a higher expected task diversity as similar tasks are likely to also rank similarly in terms of importance. There are arguments both for (more information per task) and against (harder to learn shared feature representation) higher task diversity. We therefore evaluate both approaches in our experiments.

2.3.2 Training Distantly Supervised Multitask Networks

A key problem when training neural networks on multiple tasks simultaneously using stochastic gradient descent is that gradients from the different tasks can interfere adversely (Teh et al., 2017; Doersch & Zisserman, 2017). We therefore completely disentangle the training of the unsupervised and supervised tasks in DSMT-Nets. Instead of training the auxiliary tasks jointly with the main task, we alternate between optimising DSMT-Nets for the auxiliary tasks and the main task in each epoch, starting with the auxiliary tasks. At the computational cost of an additional pass during training, the two-step training procedure prevents any potential adverse intra-step gradient interactions between the two classes of tasks. To ensure similar convergence rates for both the main and auxiliary tasks, we weight the auxiliary tasks such that the total learning rate for the unsupervised and supervised step are approximately the same, i.e. a weight of $\frac{1}{k}$ for each auxiliary task when there are k auxiliary tasks. A similar training schedule, where generator and discriminator networks are trained one after another in each iteration, has been proposed to train generative adversarial networks (GANs) (Goodfellow et al., 2014).

2.4 Experiments

We performed extensive quantitative experiments on real-world ICU data using a multitude of different hyperparameter settings in order to answer the following questions:

(1) How do DSMT-Nets perform in terms of predictive performance and label efficiency in multivariate false alarm detection relative to state-of-the-art methods for semi-supervised learning?

(2) What is the relationship between the number of auxiliary tasks, predictive performance and label efficiency?

(3) What is the importance of the architectural separation of auxiliary tasks and the main task and the two-step training procedure in DSMT-Nets?

(4) Is there value in selecting a specific set of related auxiliary tasks for distantly supervised multitask learning over random selection?

To answer question (1), we systematically evaluated DSMT-Nets and several baseline models in terms of their area under the receiver operator curve (AUC) using varying amounts of manually classified labels $n_{\text{labels}} = (12, 25, 50, 100, 500, 1244)$ and varying amounts of auxiliary tasks $k = (6, 12, 25, 50, 100)$ in the DSMT-Nets. We chose the label subsets at random without stratification. The comparison between the models' performances when using different levels of labels enable us to judge the label efficiency of the compared models, i.e. which level of predictive performance they can achieve with a limited amount of labels. By also changing the amount of auxiliary tasks used in the models, we are additionally able to assess the relationship of the number of auxiliary tasks with label efficiency and predictive performance (question (2)).

To answer question (3), we performed an ablation study using the DSMT-Nets with 100 auxiliary tasks (DSMT-Net-100) using varying amounts of manually classified labels as base models. We then trained the same models without the two-step training procedure (- two step train). In addition, we evaluated the performance of a deep Highway Network (Srivastava et al., 2015) with the same 100 auxiliary tasks distributed sequentially among layers (DSMT-Net-100D) to compare multitask learning in depth against width. Lastly, we also evaluated a multitask network where the same 100 auxiliary tasks are placed directly on the head block (Naïve Multitask Network). Through this process, we aimed to determine the relative

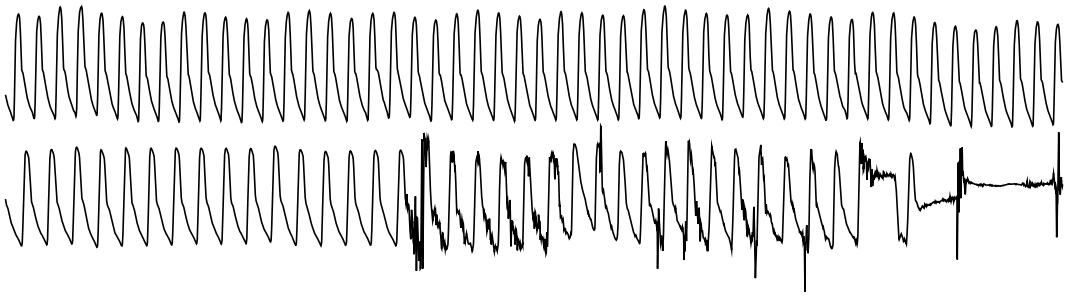


Figure 2.3: Two clear examples of arterial blood pressure signals without (top) and with pronounced artefacts (bottom). Note the high frequency noise and atypical shape in the artefact sample.

importance of the individual design choices introduced in section 2.3.

To answer question (4), we compared the predictive performance of DSMT-Nets using a random selection of all the significant features as determined by our feature selection methodology to that of DSMT-Nets that use a selection in order of feature importance. We do so with DSMT-Nets with 6 (DSMT-Net-6R) and 100 (DSMT-Net-100R) auxiliary tasks to additionally assess whether the importance of auxiliary task selection is sensitive to the number of auxiliary tasks.

In total, we trained 2 730 distinct model configurations in order to gain a better understanding of the empirical strengths and weaknesses of DSMT-Nets.

2.4.1 Dataset

We collected biosignal monitoring data from January to August 2017 (8 months) from consenting patients admitted to the Neurocritical Care Unit at the University of Zurich, Switzerland. The data included continuous, evenly-sampled waveforms obtained by electrocardiography (ECG; 200 Hz), arterial blood pressure (ART; 100 Hz), pulse oximetry (PPG and SpO_2; 100 Hz) and intracranial pressure (ICP; 100 Hz) measurements. For this study, we did not collect or make use of any personal, demographic or clinical data, such as prior diagnoses, treatments or electronic health records.

To obtain a ground truth assessment of alarms, we provided clinical staff with a user interface and instructions[2] for annotating alarms that they believed were caused by artefacts or a technical error (Figure 2.3). Because of technical limitations in exporting data from the biosignal database, we selected the subset of 20 monitoring days of 14 patients with the highest amount of manually labelled alarms for further analysis. The evaluated dataset encompassed a grand total of 13 938 alarms, yielding an average rate of 696.9 alarms per patient per day. This number is in line with alarm rates reported in previous works (Cvach, 2012). Of all alarms, 46.99% were caused by an alarm-generating algorithm operating on the ART waveform, 33.10% on PPG or derived SpO_2, 12.02% on ICP and 7.89% on either ECG or an ECG-derived signal, such as the heart rate signal.

Annotations. Out of the whole set of alarms, 1 777 (12.75%) alarms were manually labelled by clinical staff during the observed period. Because we used multiple annotators that were not calibrated to each other's assessments, we additionally conducted a review over all 1 777 annotations in order to ensure the internal consistency of the set of annotations as a whole. In this review round, we found that a total of 603 (33.93%) annotations were inconsistent. We subsequently assigned corrected labels to these alarms. Since label quality is paramount for model training and validation, we suggest at least one label review round using the majority vote of a committee of labellers with clear instructions in order to maintain a sufficient degree of label consistency. Recent large-scale labelling efforts in physiological monitoring of arrhythmias (Clifford et al., 2017) suggest that even more review rounds might be necessary to obtain a gold standard set of labels. In our final label set, 976 (45.08%) out of all annotated alarms are labelled as most likely being caused by an artefact or technical error. We note that a data collection effort of this scale is extremely expensive and therefore economically infeasible for most hospitals, motivating our search for a more label-efficient approach.

[2]Detailed instructions and full qualitative samples can be found in the supplementary material.

2.4.2 Evaluation Setup

As input data, we extracted a 40 second window of the time frame immediately before an alarm was triggered from each available biosignal. We considered a signal stream to be missing for a given alarm setting if the last recorded measurement of that type happened longer than 10 seconds ago. To reduce the computational resources required for our experiments, we resampled the input data to $\frac{1}{16}^{th}$ of its original sampling rate. In our preliminary evaluation, we did not see significant performance changes when using a higher sampling rate or a longer context window. Additionally, we normalised the extracted windows of each stream to the range of $[-1, 1]$ using the maximum and minimum values encountered in that window.

Baselines. To ensure a fair reference, we used the DSMT-Nets base architecture without any horizontal blocks and auxiliary tasks as the supervised baseline (DSMT-Net-0, Figure 2.2a). Because the supervised baselines have no auxiliary tasks, we trained them in a purely supervised manner on the labelled alarms only.

As a baseline for feature selection, we used the automated feature extraction and selection approach from (Christ et al., 2016) to identify a large number (up to 875) of relevant time series features from the multivariate input data. Note that we followed this process separately for each distinct amount of labels in order to avoid information leakage. We then fed those features to a random forest (RF) classifier consisting of 4 096 trees to produce predictions (Feature RF). As mentioned in section 2.3.2, we used the same feature selection approach to identify suitable auxiliary tasks for DSMT-Nets. The Feature RF baseline therefore serves as a reference for directly using the identified significant features to make a prediction.

For comparison to the state-of-the-art in reconstruction-based semi-supervised learning, we evaluated Ladder Networks (Rasmus et al., 2015) on the same dataset. We replaced the DSMT-Net components on top of the joint feature representation P_c with a Ladder Network in order to

use a comparable architecture that is also able to model missingness and heterogenous input streams.

For comparison to the state-of-the-art in semi-supervised adversarial learning, we trained GANs using a semi-supervised objective function and feature matching (Salimans et al., 2016) on the same dataset. This type of GAN has been shown to be highly efficacious at semi-supervised learning in low-resolution image datasets (Salimans et al., 2016). We trained the generator networks to generate a context window of multiple high-resolution time series as input to a DSMT-Net discriminator without any auxiliary tasks. In terms of architecture, the generator networks used strided upsampling convolutions.

Hyperparameters. To ensure a fair comparison, we used a systematic approach to hyperparameter selection for each evaluated neural network. We trained each model 35 times with a random choice of the three variable hyperparameters bound to the same ranges ($1 - 3$ hidden layers, $16 - 32$ units/filters per hidden layer, $25\% - 85\%$ dropout). We reset the random seed to the same value for each model in order to make the search deterministic across training runs, i.e. all the models were evaluated on exactly the same set of hyperparameter values. Note that this setup does not guarantee optimality for any model, however, with respect to the evaluated hyperparameters, it guarantees the models were evaluated fairly and given the same amount of scrutiny. To train the neural network models, we used a learning rate of 0.001 for the first ten epochs and 0.0001 afterwards to optimise the binary cross-entropy for the main classification output and the mean squared error for all auxiliary tasks. We additionally used early stopping with a patience of 13 epochs. For the extra hyperparameters in Ladder Networks, we set the noise level to be fixed at 0.2 at every layer, the denoising loss weight to 100 for the first hidden layer and to 0.1 for every following hidden layer. For the GAN models, we used a base learning rate of 0.0003 for the discriminator and a slightly increased learning rate of 0.003 for the generator to counteract the faster convergence of the discriminator networks. We trained GANs using an early stopping patience on the main loss of 650 steps for a minimum of 2 500 steps. To choose these extra

hyperparameters of GANs and Ladder Networks, we followed the original author's published configurations (Rasmus et al., 2015; Salimans et al., 2016) and adjusted them slightly to ensure they converged.

Architectures. We used the conceptual architecture from Figure 2.2 as a base architecture for the DSMT-Nets. As perception blocks, we employed ResNets (He et al., 2016) with 1-dimensional convolutions over the time axis for each input data stream. As head block and multitask blocks, we used Highway Networks (Srivastava et al., 2015). The head block hosted a sigmoid binary output y that indicated whether or not the proposed alarm was likely caused by an artefact. In addition, we used batch normalisation in the DSMT-Net blocks.

Metrics. For each approach, we report the AUC of the best model encountered over all 35 hyperparameter runs.

Dataset Split. We applied a random split stratified by alarm classification to the whole set of annotated alarms to separate the available data into a training (70%, 1 244 alarms) and test set (30%, 533 alarms).

2.5 Results and Discussion

We report the results of our experiments in Table 2.1 and discuss them in the following paragraphs.

Predictive Performance. Overall, we found that the label limit after which the purely supervised approaches consistently outperformed the semi-supervised approaches was between 100 and 500 labels. The strongest approach when using all 1 244 available labels and the 500 label subset was the purely supervised Feature RF baseline. Out of all compared methods, DSMT-Nets were the most label-efficient approach when using 25, 50 and 100 labels. However, the Feature Matching GAN outperformed the DSMT-Nets when using just 12 labels. In our experimental setting, the best DSMT-Nets yielded significant improvements in AUC over both reconstruction-based as well as adversarial state-of-the-art approaches to semi-supervised

Table 2.1: Comparison of the maximum AUC value across the 35 distinct models (vertical) that we trained using different sets of hyperparameters and varying amounts of labels (horizontal). We report the AUC of the best encountered model as calculated on the test set of 533 alarms. The best results in each column are highlighted in bold. The standard deviation and minimum observed AUC value across the 35 models trained using different hyperparameters are given in Tables 2.2 and 2.3, respectively.

AUC with # of Labels	12	25	50	100	500	1 244
Feature RF	0.567	0.574	0.628	0.822	**0.942**	**0.955**
Supervised baseline	0.751	0.753	0.806	0.873	0.941	0.942
Naïve Multitask Network	0.791	0.804	0.828	0.887	0.941	0.940
Ladder Network	0.791	0.772	0.800	0.842	0.863	0.868
Feature Matching GAN	**0.846**	0.834	0.834	0.865	0.911	0.898
DSMT-Net-6	0.763	0.839	0.866	0.897	0.924	0.934
DSMT-Net-12	0.739	**0.872**	0.891	0.890	0.928	0.933
DSMT-Net-25	0.761	0.870	0.886	0.898	0.924	0.929
DSMT-Net-50	0.722	0.847	**0.901**	0.906	0.926	0.936
DSMT-Net-100	0.720	0.831	0.893	0.907	0.934	0.934
- two step train	0.733	0.798	0.785	0.814	0.849	0.898
DSMT-Net-6R	0.805	0.851	0.884	**0.909**	0.921	0.938
DSMT-Net-100R	0.790	0.860	0.883	**0.909**	0.918	0.932
DSMT-Net-100D	0.587	0.611	0.722	0.610	0.624	0.702

learning on low-resolution image benchmarks. The relative improvements in AUC amounted to 13.0%, 12.6% and 8.0% over Ladder Networks and 9.4%, 10.4% and 5.6% over Feature Matching GANs at 25, 50 and 100 labels, respectively. We note that even Naïve Multitask Networks, that did not make use of any of the adaptions introduced by DSMT-Nets, with the exception of two cases outperformed both Ladder Networks and Feature Matching GANs - suggesting that distant supervision in general is an efficacious approach to semi-supervised learning in this domain.

Interestingly, most of the evaluated semi-supervised approaches, with the exception of Naïve Multitask Networks, were outperformed by their purely supervised counterparts at lower amounts of labels than one would expect - in many cases by a large margin. Indeed, both Feature Matching GANs as well as Ladder Networks were eclipsed by the supervised baseline at just 100 labels. This suggests that either: (i) Feature Matching GANs and Ladder Networks require a higher degree of hyperparameter optimisation than the other evaluated approaches or (ii) the strengths of these approaches in the domain of low-resolution images do not generalise to the same degree to the domain of multivariate high-resolution time series without adaptations. These are novel findings given that most other recent evaluations of state-of-the-art methods in semi-supervised learning have been confined solely to the low-resolution image domain. We believe that, in the future, more systematic replication studies, such as the one presented in this work, are necessary to evaluate the degree to which new methods generalise beyond benchmark datasets that often do not cover many practically important data modalities, such as time series data, and idiosyncrasies, such as missingness, heterogeneity, sparsity and noise.

In terms of sensitivity and specificity, our best models would have been able to reduce the number of false alarms brought to the attention of clinical staff with the same degree of urgency as true alarms with sensitivities of 22.97% (Feature Matching GAN), 40.99% (DSMT-Net-12), 48.76% (DSMT-Net-50), 63.60% (DSMT-Net-100R), 66.43% (Feature RF) and 76.68% (Feature RF) using, respectively, 12, 25, 50, 100, 500 and 1 244 labelled training samples at a specificity of 95%. In relative terms, DSMT-Nets were therefore

- with just 100 labels - able to realise $\frac{63.60}{76.68} = 82.94\%$ of the expected reduction in false alarms of the Feature RF that was trained on 1 244 labels. This finding confirms that a modest data collection effort would be sufficient to achieve a considerable improvement in false alarm rates in critical care.

Number of Auxiliary Tasks. In DSMT-Nets with auxiliary tasks selected by feature importance, more auxiliary tasks achieved slightly better performances once sufficient amounts of labels were available. We reason that, because the head block was trained on labelled samples only, a greater number of labels was necessary to effectively orchestrate the extra information provided by a larger number of multitask blocks. However, we did not see the same behavior in DSMT-Nets with auxiliary tasks selected at random. Here, the performances of DSMT-Nets with 6 and 100 auxiliary tasks were comparable across all label levels.

Importance of Adaptions. We found that using DSMT-Nets trained with auxiliary tasks distributed in depth (DSMT-Net-100D) performed worse than our proposed architecture - demonstrating that parallel alignment of multitask blocks is the superior architectural design choice. Similarly, DSMT-Net-100 variants without the two step training procedure (- two step train) consistently failed to reach the semi-supervised performance of their counterparts with the two step training procedure enabled (DSMT-Net-100) for more than 12 labels. This shows that disentangling the training of the auxiliary and the main task played an integral role in the strong semi-supervised performance of DSMT-Nets and further reinforces prior reports that adverse gradient interactions are a key challenge for multitask learning in neural networks (Teh et al., 2017; Doersch & Zisserman, 2017).

Task Selection. We found that random selection in most cases outperformed selection in order of feature importance when comparing the DSMT-Net-6 and DSMT-Net-6R variants. We believe this was the result of increased task diversity when selecting at random from the relevant auxiliary tasks, as similar features rank close to each other in terms of feature importance. The fact that this effect was less pronounced between the same models with more auxiliary tasks (DSMT-Net-100R and DSMT-

Net-100) supports this theory, as a larger set of tasks will automatically have a higher diversity due to the limited amount of highly similar features, thus decreasing the importance of accounting for diversity in the selection methodology. We therefore conclude that task diversity is the dominant factor in selecting related auxiliary tasks for distant multitask supervision.

2.6 Limitations

False alarms in the ICU are not solely a technical problem (Cvach, 2012; Drew et al., 2014). Organisational and processual aspects must also be considered to comprehensibly address this issue in clinical care (Drew et al., 2014). One such aspect is the question of how to best manage those alarms that have been flagged as false by an alarm classification system. We reason that, due to the inherent possibility of suppressing a true alarm, a sensible approach would be to report those errors with a lower degree of urgency, i.e. with a less pronounced sound, rather than completely suppressing them (Cvach, 2012).

Another limitation of this work is that we only considered the detection of alarms that are caused by either artefacts or technical errors. Alarms that are technically correct, but clinically require no intervention, are another important source of false alarms (Drew et al., 2014) that we did not analyse in this work. Identifying clinically false alarms is significantly harder than those caused by artefacts and technical errors, as clinical reasoning requires deep knowledge of a patient's high-level physiological state, as well as a significant amount of domain knowledge.

Lastly, while the presented distantly supervised approach to semi-supervised learning performs well on our dataset, its applicability to other datasets hinges on being able to determine multiple related auxiliary tasks. We only evaluated distantly supervised multitask learning on time series data, where large numbers of suitable auxiliary tasks are readily available through automated feature extraction and selection (Christ et al., 2016).

We hypothesise that it might not be trivial to find large repositories of auxiliary tasks suitable for distant multitask supervision for all data types. A comparatively small number of potential auxiliary tasks have been reported in related works in computer vision and natural language processing (Blaschko et al., 2010; Xu et al., 2015a; Oquab et al., 2015; Deriu et al., 2017; Doersch & Zisserman, 2017). Finally, our experiments yield insights into the importance of auxiliary task selection in DSMT-Nets, but further theoretical analyses are necessary to understand exactly what types of auxiliary task are useful to what degree in distantly supervised multitask learning.

2.7 Conclusion

We presented a novel approach to reducing false alarms in the ICU using data obtained from a dynamic set of multiple heterogenous biosignal monitors. Unlabelled data is abundantly available, but obtaining trustworthy expert labels is laborious and expensive in this setting. We introduced a multitask network architecture that leverages distant supervision through multiple related auxiliary tasks in order to reduce the number of expensive labels required for training. We developed both a methodology for automatically selecting auxiliary tasks from multivariate time series as well as an optimised training procedure that counteracts adverse gradient interactions between tasks. Using a real-world critical care dataset, we demonstrated that our approach leads to significant improvements over several state-of-the-art baselines. In addition, we found that task diversity and adverse gradient interactions are key concerns in distantly supervised multitask learning. Going forward, we believe that our approach could be applicable to a wide variety of machine-learning tasks in healthcare for which obtaining labelled data is a major challenge.

2.8 Supplementary Material

2.8.A Source Code

The source code for this work is available online at
`https://github.com/d909b/DSMT-Nets`.

2.8.B Instructions for Annotators

We instructed our annotators to label a given alarm context window as caused by an artefact if:

1. The signal that caused the alarm is not being recorded, as verified by visibility on the monitor.
2. The alarm-generating signal curve has an atypical shape.
3. Numerical values derived from the alarm-generating signal are not physiologically plausible.

Figures 2.4 and 2.5 depict qualitative examples of context windows that have been labelled as caused by an artefact.

Table 2.2: Comparison of the standard deviation of AUC values across the 35 distinct models (vertical) that we trained using different sets of hyperparameters and varying amounts of labels (horizontal). We report the AUC of the best encountered model as calculated on the test set of 533 alarms. The worst result in each column is highlighted in bold. A higher variation in AUC across hyperparameter choices and training runs may indicate higher sensitivity to hyperparameters in the evaluated range and/or lacking robustness of training in the presented setting. Most notably, we find that disentangling training of the auxiliary and the main task in DSMT-Nets improves training stability in most cases.

AUC with # of Labels	12	25	50	100	500	1 244
Feature RF	-	-	-	-	-	-
Supervised baseline	0.055	0.046	0.045	0.026	0.008	0.007
Naïve Multitask Network	0.061	0.057	0.048	0.054	0.049	0.041
Ladder Network	0.067	0.074	0.069	0.076	**0.066**	**0.076**
Feature Matching GAN	0.050	0.051	0.051	0.037	0.020	0.027
DSMT-Net-6	0.059	0.058	0.056	0.070	0.040	0.041
DSMT-Net-12	0.058	0.064	0.072	0.074	0.040	0.037
DSMT-Net-25	0.066	0.062	0.068	0.076	0.038	0.043
DSMT-Net-50	0.059	0.071	0.066	0.060	0.042	0.044
DSMT-Net-100	0.060	0.060	**0.075**	0.048	0.032	0.039
- two step train	**0.070**	**0.076**	0.065	**0.078**	0.058	0.061
DSMT-Net-6R	0.058	0.051	0.061	0.062	0.039	0.035
DSMT-Net-100R	0.070	0.062	0.054	0.047	0.034	0.048
DSMT-Net-100D	0.019	0.023	0.038	0.021	0.030	0.038

Table 2.3: Comparison of the minimum AUC value across the 35 distinct models (vertical) that we trained using different sets of hyperparameters and varying amounts of labels (horizontal). We report the AUC of the best encountered model as calculated on the test set of 533 alarms. The best results in each column are highlighted in bold. The difference between the maximum and minimum value indicates the range of values covered over the 35 hyperparameter settings.

AUC with # of Labels	12	25	50	100	500	1 244
Feature RF	-	-	-	-	-	-
Supervised baseline	0.501	0.547	0.568	**0.763**	**0.907**	**0.911**
Naïve Multitask Network	0.516	0.577	0.613	0.648	0.693	0.732
Ladder Network	0.506	0.516	0.538	0.512	0.594	0.560
Feature Matching GAN	**0.629**	**0.628**	0.646	0.719	0.817	0.757
DSMT-Net-6	0.514	0.557	0.588	0.604	0.760	0.752
DSMT-Net-12	0.507	0.540	0.579	0.630	0.753	0.791
DSMT-Net-25	0.501	0.603	0.535	0.570	0.774	0.779
DSMT-Net-50	0.506	0.557	0.649	0.682	0.768	0.770
DSMT-Net-100	0.507	0.552	0.600	0.691	0.797	0.774
- two step train	0.502	0.500	0.539	0.525	0.645	0.685
DSMT-Net-6R	0.515	0.624	0.630	0.635	0.760	0.805
DSMT-Net-100R	0.506	0.601	**0.660**	0.686	0.771	0.771
DSMT-Net-100D	0.500	0.500	0.500	0.500	0.500	0.500

Table 2.4: The exact hyperparameter values used for each model for each of the 35 distinct training runs. We chose the values using a uniformly random selection within the ranges specified in the main paper. The number of hidden units per layer and the number of hidden layers were rounded to the nearest integer in our experiments.

Run	Dropout	Number of hidden units / layer	Number of hidden layers
1	0.5256	18.3015	1.4562
2	0.2926	26.3799	1.6650
3	0.3888	29.7946	1.4185
4	0.4633	29.1221	1.7195
5	0.3619	27.6884	1.7030
6	0.5049	26.5369	2.7647
7	0.7134	26.2866	1.4111
8	0.4486	23.7360	2.4363
9	0.2939	24.0741	1.1734
10	0.5652	21.2195	1.2685
11	0.3688	18.8924	2.5907
12	0.7542	20.2902	2.7300
13	0.2614	27.6143	1.5102
14	0.3820	24.7860	2.1281
15	0.3452	25.3250	2.9806
16	0.7308	30.3649	1.4315
17	0.6195	22.6811	1.7044
18	0.6170	21.3986	2.7229
19	0.7451	27.8114	2.2333
20	0.3469	22.9611	1.4900
21	0.5168	16.2036	2.9124
22	0.4098	20.5713	2.4480
23	0.3012	24.5169	1.3481
24	0.4475	17.3175	2.8138
25	0.2660	27.0517	1.2606
26	0.4830	21.8282	2.9766
27	0.7799	18.0746	2.1824
28	0.3712	24.3822	2.1989
29	0.5958	25.3871	2.8844
30	0.2649	30.3633	2.6249
31	0.6065	20.6158	1.9874
32	0.4623	16.1852	1.3220
33	0.2592	24.9682	1.8996
34	0.6531	26.4506	2.3409
35	0.7825	28.5137	2.9273

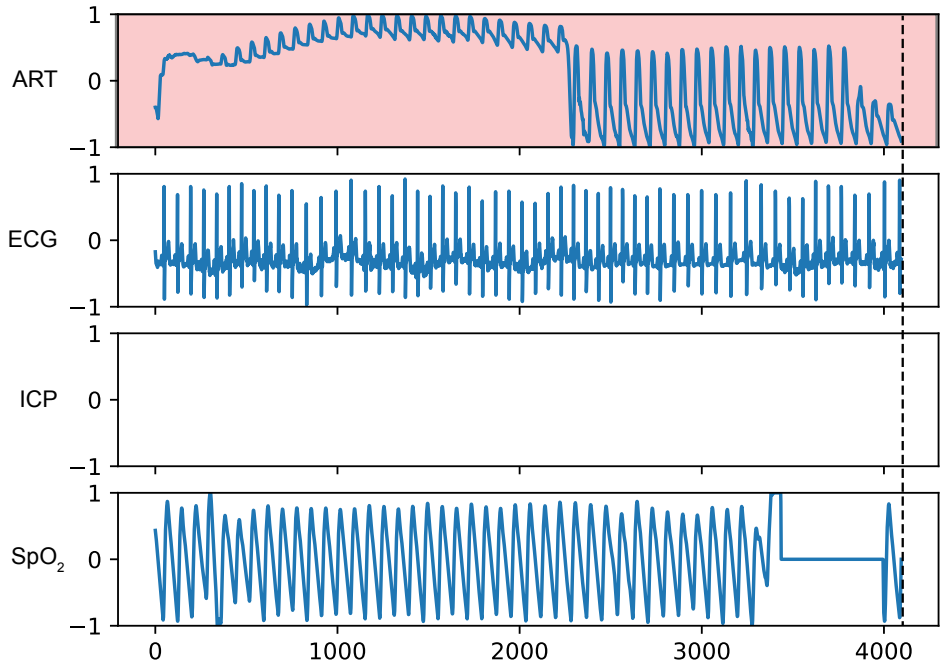


Figure 2.4: A qualitative example of an alarm caused by an artefact, as encountered in the ICU dataset. Depicted are the amplitudes (y-axis, normalised) over time (x-axis, in hundredths of a second) of the arterial blood pressure (ART), electrocardiography (ECG), intracranial pressure (ICP) and pulse oximetry (SpO_2) signals immediately before the alarm was triggered. The dashed line spanning all biosignals indicates the time at which an alarm was triggered. An empty box indicates a missing signal. In this case, the alarm was triggered by the arterial blood pressure monitor (red). Note that there also appears to be an artefact in the pulse oximetry signal that might have triggered another independent alarm concurrently.

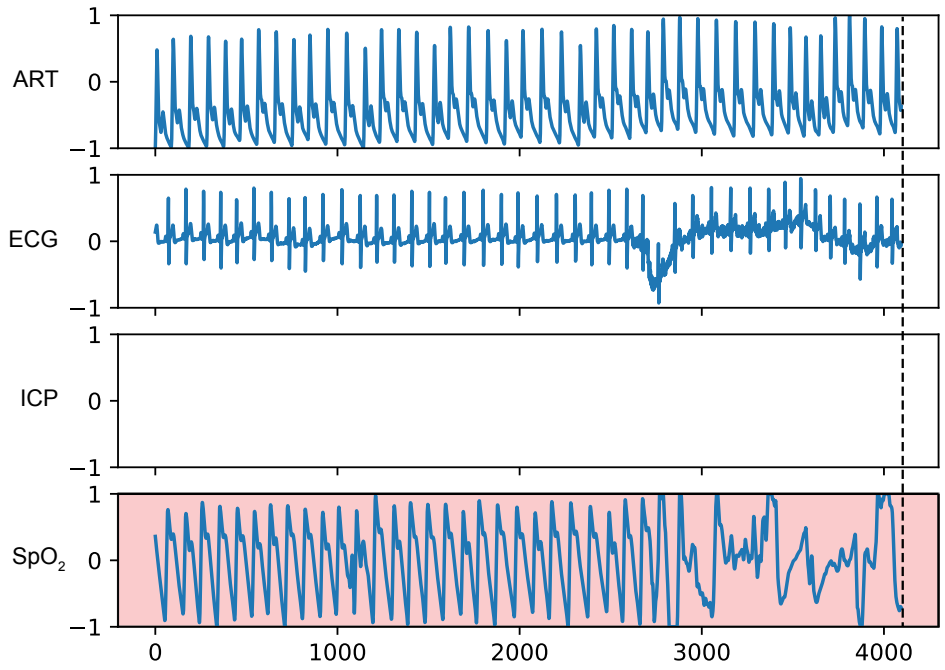


Figure 2.5: A qualitative example of an alarm caused by an artefact, as encountered in the ICU dataset. Depicted are the amplitudes (y-axis, normalised) over time (x-axis, in hundredths of a second) of the arterial blood pressure (ART), electrocardiography (ECG), intracranial pressure (ICP) and pulse oximetry (SpO_2) signals immediately before the alarm was triggered. The dashed line spanning all biosignals indicates the time at which an alarm was triggered. An empty box indicates a missing signal. In this case, the alarm was triggered by the pulse oximetry monitor (red).

This page was intentionally left blank.

Chapter 3

Learning to Diagnose Parkinson's Disease from Smartphone Data

"You don't understand anything until you learn it more than one way." - Marvin Minsky

Parkinson's disease is a neurodegenerative disease that can affect a person's movement, speech, dexterity, and cognition. Clinicians primarily diagnose Parkinson's disease by performing a clinical assessment of symptoms. However, misdiagnoses are common. One factor that contributes to misdiagnoses is that the symptoms of Parkinson's disease may not be prominent at the time the clinical assessment is performed. Here, we present a machine-learning approach towards distinguishing between people with and without Parkinson's disease using long-term data from smartphone-based walking, voice, tapping and memory tests. We demonstrate that our attentive deep-learning models achieve significant improvements in predictive performance over strong baselines (area under the receiver operating characteristic curve = 0.85) in data from a cohort of 1 853 participants. We also show that our models identify meaningful features in the input data. Our results confirm that smartphone data collected over extended periods of time could in the future potentially be used as a digital biomarker for the diagnosis of Parkinson's disease.

3.1 Introduction

Parkinson's disease (PD) affects more than 6 million people worldwide (Vos et al., 2016) and is the second most common neurodegenerative disease

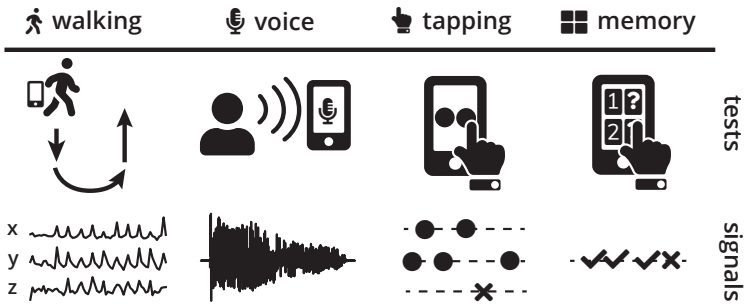


Figure 3.1: Smartphones can be used to perform tests that are designed to trigger symptoms of Parkinson's disease (top). During these tests, smartphone sensors record high-resolution signals (bottom) that we can use to distinguish between people with and without Parkinson's disease.

after Alzheimer's disease (De Lau & Breteler, 2006). The symptoms of PD progressively worsen over time, leading to a stark loss in quality of life (Schrag et al., 2000), and a significant reduction in life expectancy (De Lau & Breteler, 2006). While there currently exists no cure for PD, available pharmacological and surgical treatment options are effective at managing the symptoms of PD (Goetz et al., 2005; Connolly & Lang, 2014). Receiving a timely and accurate diagnosis is paramount for patients because access to treatments could improve their quality of life (Global Parkinson's Disease Survey Steering Committee, 2002). Currently, clinicians diagnose PD primarily based on subjective clinical assessments of patients' symptoms (Pahwa & Lyons, 2010). However, research has shown that around 25% of PD diagnoses are incorrect when compared to results of post-mortem autopsy (Pahwa & Lyons, 2010). Diagnosing PD is difficult because there are other movement disorders that may appear similar to PD, and because symptom severity in PD may fluctuate over time (Pahwa & Lyons, 2010).

Smartphone-based tests could potentially give clinicians access to long-term measurements of symptom severity and symptom fluctuation by enabling patients to record themselves outside the clinic (Figure 3.1). However, making sense of observational smartphone data is extremely challenging for both humans and machines due to the large number of diverse data streams sampled at high resolution over long periods of

time. Major unsolved questions include how to simultaneously cover the wide range of symptoms associated with PD, how to best aggregate the vast amounts of clinically relevant data collected over time, and how to communicate the decisions of predictive models to clinicians.

To address these issues, we present a novel approach towards distinguishing between people with and without PD from smartphone data. Our method is built on the idea of first training specialised models to assess symptom severity from single test instances, and then using an evidence aggregation model to aggregate an arbitrary number of assessments from several types of tests into a final prediction. We extend our method with hierarchical attention to visualise both the importance of tests as well as the importance of segments in those tests towards a prediction. Our experiments demonstrate that this approach leads to significant improvements in predictive performance over several strong baselines, and highlight the potential of smartphones to become accessible tools for gathering clinically relevant data in the wild.

Contributions. Our contributions in this chapter are as follows:

- We present machine-learning models to assess symptoms of PD from signals recorded during smartphone-based walking, voice, tapping and memory tests.
- We introduce an evidence aggregation model (EAM) to integrate arbitrary numbers of symptom assessments from multiple types of tests over long periods of time to produce a single diagnostic score.
- We develop a hierarchical neural attention mechanism that quantifies the importance of both individual tests and segments within those tests towards the diagnostic score.
- We perform experiments on real-world data collected from 1 853 mPower participants with and without PD that show that our approach leads to significant improvements in prediction performance over several strong baselines.

3.2 Related Work

Background. Machine learning has a rich history in facilitating medical diagnoses. Machine learning has, for example, been applied to diagnosing breast cancer from tumor features (Zheng et al., 2014), cardiac arrhythmias and cardiac risk factors from smartphone-based heart rate sensors (Oresko et al., 2010; Schwab et al., 2017; Ballinger et al., 2018), skin cancer from clinical images (Esteva et al., 2017), depressed moods from information self-reported via smartphones (Suhara et al., 2017), and a wide range of clinical diagnosis codes from electronic health records and lab test results (Lipton et al., 2016a; Choi et al., 2016a; Razavian et al., 2016). Predicting a person's disease status is difficult because there is a vast range of factors that may influence an individual's health. Wearable sensors and smart devices enable us to capture a number of these factors with minimal burden on users by passively and continuously tracking behaviors and environmental factors (Quisel et al., 2017). However, in contrast to clean, standardised benchmark datasets, observational data collected by wearable sensors and smart devices in the real-world is often difficult to integrate with existing machine-learning approaches. The difficulty of applying existing machine-learning methods to complex datasets has led to the development of specialised methods to deal with several of the idiosyncrasies of observational health data, such as missingness (Lipton et al., 2016b; Che et al., 2018), long-term temporal dependencies (Choi et al., 2016a), noise (Schwab et al., 2017), heterogeneity (Libbrecht & Noble, 2015), irregular sampling (Lipton et al., 2016a), sparsity (Lasko et al., 2013), and multivariate input data (Ghassemi et al., 2015; Schwab et al., 2018a). However, adapting existing machine-learning methods to account for the idiosyncrasies of healthcare data remains an ongoing challenge (Ghassemi et al., 2018).

Monitoring and Diagnosis of PD. There has been much interest in leveraging new technologies and data modalities to better diagnose and assess symptom severity in PD. There are a number of driving factors behind the interest in new approaches: Firstly, despite the severity of the disease,

clinical PD diagnoses are currently relatively inaccurate. Diagnoses are particularly difficult in the earlier stages of the disease and in the presence of other disorders that may appear similar to PD (Rizzo et al., 2016). Secondly, new technologies could lead to patients receiving their diagnoses earlier. An early diagnosis could potentially improve a patient's quality of life by giving them access to symptom-suppressing treatments (Global Parkinson's Disease Survey Steering Committee, 2002). Lastly, both clinical trials for new pharmaceutical treatments and clinical decision-making require the ability to accurately diagnose and objectively assess symptoms of PD (Shulman et al., 2006; Dorsey et al., 2017). Previous works have for example used data from pen movements (Smith et al., 2007), wearables (Patel et al., 2009; Klucken et al., 2013), and speech features (Little et al., 2007, 2009; Tsanas et al., 2010, 2011, 2012) to objectively monitor or diagnose PD. A number of works have also proposed the use of smartphone sensors for continuously monitoring symptoms in PD (Hammerla et al., 2015; Arora et al., 2015; Zhan et al., 2016, 2018; Prince et al., 2018). Recently, the PD Digital Biomarker DREAM challenge[1] aimed to develop machine-learning models to diagnose PD from accelerometer data in a collaborative effort. (Emrani et al., 2017) proposed a multitask-learning framework to identify biomarkers that are predictive of progression in PD. However, their approach did not integrate raw sensor data and could not handle missing input data.

In contrast to existing works, we present the first machine-learning approach to distinguishing between people with and without PD that integrates information from sensor measurements of several types of smartphone-based tests over long periods of time. Our approach is able to simultaneously (i) assess single test instances and (ii) produce a unified diagnostic score. In addition, we introduce a hierarchical neural attention mechanism that enables us to reason about both the importance of specific tests as well as the importance of individual segments within those tests towards the final diagnostic score. Furthermore, we perform our experiments on data collected from 1 853 mPower participants, the largest cohort used to validate a machine-learning approach to diagnosing PD from

[1] `http://synapse.org/DigitalBiomarkerChallenge`; accessed 1st September 2018

smartphone data to date.

3.3 Methodology

Overview. We utilise data collected during the mPower study, a large-scale observational study about PD conducted entirely through a smartphone app (Bot et al., 2016). In the study, participants with and without PD are asked to perform four smartphone-based tests (walking, voice, tapping and memory; Figure 3.1) up to three times a day without any supervision. In addition to regularly performing the tests, participants provide their detailed demographic profile, including possible prior clinical diagnoses of PD, using self-reporting forms within the app[2]. The main idea of the presented approach is to connect the sensor data collected by the participants' smartphones with their prior professional diagnoses to train machine-learning models to learn to diagnose PD.

Smartphone Tests. mPower participants perform the following four types of tests using their personal smartphones (Bot et al., 2016):

- **Walking Test.** To perform the walking test, participants are asked to put their smartphone in their pocket, walk 20 steps forward, turn around, stand still for 30 seconds, and then walk 20 steps back. We denote the three distinct segments of the walking test as: Outbound, rest, and return, respectively. During the test, the smartphone's accelerometer and gyroscope record the participant's three-dimensional linear and angular acceleration. This test is designed to measure movement impairments associated with PD, such as tremor, rigidity, and freezing of gait.

- **Voice Test.** In the voice test, participants are asked to say "aaaah" into their smartphones' microphone for up to 10 seconds. The smartphone's microphone records the audio data during the test and during the preceding countdown. The goal of the audio test is to

[2] `https://parkinsonmpower.org/`; accessed 1st September 2018

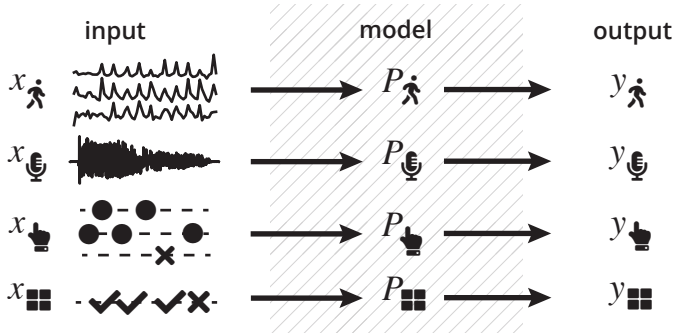


Figure 3.2: An illustration of the data processing pipelines for each of the test types (vertical) from the input signals $x_\star$ (left) over the specialised predictive models $P_\star$ (center) to the single-test output predictions $y_\star$ (right). The use of specialised predictive models for each test type enables us to choose the most suitable model for each of the heterogenous input signals.

expose speech impairments that are commonly found in people with PD.

- **Tapping Test.** In the tapping test, participants are asked to position their smartphones on a flat surface and alternatingly tap two buttons on the screen for 20 seconds. The smartphone records the positions and timestamps of the participant's taps on the screen. In addition, the smartphone's accelerometer measures the three-dimensional movements of the smartphone during the test. The tapping test is aimed at uncovering signs of impaired finger dexterity. Impaired finger dexterity is a common symptom in people with PD.

- **Memory Test.** In the memory test, participants are presented with a grid of flowers on their smartphone screens. During the test, different flowers are illuminated one at a time. Participants are then asked to repeat the observed sequence by touching the flowers in the same order. The collected data includes the positions and timestamps of the participant's taps on the smartphone's screen and the sequence order as displayed to the participant. This test measures the spatial memory of the participant, which may be impaired due to PD (Bot et al., 2016).

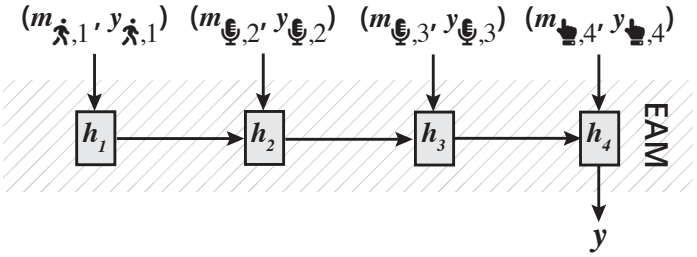


Figure 3.3: Temporal ensembling using an evidence aggregation model (EAM). The EAM (grey) receives per-test metadata ($m_{\star,j}$) and per-test output predictions ($y_{\star,j}$) in temporal order as input. In this example, the EAM's hidden state (h_j) aggregates the information from the $k = 4$ performed tests to produce a final output y that indicates whether or not the participant is likely to have PD.

Multistage Approach. Our approach to distinguishing between people with and without PD consists of two stages. In the first stage, we use specialised predictive models $P_\star$ to identify PD in signals $x_\star$ from a single type of test with $\star \in \{$ (walking), (voice), (tapping), (memory)$\}$. The specialised models are trained to predict a participant's diagnosis given the signal data from exactly one sample instance of one type of test (Figure 3.2). The output of the specialised models is a local prediction $y_\star$ that indicates, on a scale from 0 to 1, how likely it is that the participant that performed the given test instance has PD:

$$y_\star = P_\star(x_\star) \tag{3.1}$$

The specialised models $P_\star$ are the building blocks for the second stage. In the second stage, the outputs $y_{\star,j}$, with $j \in [1\,..\,k]$, of the specialised models and the metadata $m_{\star,j}$ for all k tests performed by a user are aggregated into a single diagnostic prediction y using an EAM (Figure 3.3):

$$y = \text{EAM}([(m_{\star,1}, y_{\star,1}), ..., (m_{\star,k}, y_{\star,k})]) \tag{3.2}$$

The primary idea behind Equations 3.1 and 3.2 is to disentangle learning how to assess symptom severity from each test and how to aggregate

multiple tests over a period of time. This compositional approach to modelling the problem of diagnosing PD from a range of diverse smartphone tests enables us to choose the most suitable predictive model for the various test types and the EAM. Furthermore, each specialised predictive model $P_\star$ is optimised for one type of test only. In contrast to an end-to-end model, the specialised predictive models do not need to consider how to aggregate multiple tests and which patterns may be important in other tests. Similarly, the EAM is entirely focused on learning how to best aggregate the evidence from multiple tests. In essence, our approach follows a divide-and-conquer approach by ensuring that each component is focused on exactly one task. Another benefit of the given abstract formulation is that it enables us to choose from a wide range of models for both the specialised predictive models and the EAM, since there are no specific requirements on either other than that they need to process $x_\star$ and tuples of $(m_{\star,i}, y_{\star,i})$, respectively.

Hierarchical Neural Attention. In addition to the diagnostic score y, our approach provides the clinician with information about which tests and test segments in the data recorded by the user were most important for the model's output. Presenting information about which data the model output is based on can help put the diagnostic score y in perspective and inform the clinician's further clinical decision-making. For example, when confronted with a patient whose diagnostic prediction focused primarily on motor symptoms, the clinician can focus her efforts on ruling out other movement disorders that may cause similar symptoms. In order to highlight (i) which individual tests were most important for the EAM's output y, and (ii) which segments of specific tests were most important for the local predictions $y_\star$, we introduce a hierarchical neural soft attention mechanism. When using neural networks as predictive models, the upper-level attention mechanism (i) is a component of the EAM and the lower-level attention mechanism (ii) is part of the specialised models $P_\star$. Both the upper- and lower-level attention mechanism use the same mathematical formulation. Given the topmost hidden feature representations h_i of (i) all the tests performed by a user, or (ii) segments in the recorded signal streams for a single test, we calculate

Table 3.1: Comparison of the AUC and AUPR values for the different test types when only given the data of a single test to make a diagnostic decision. We compared the performances of neural networks (CNN, RNN) with expert features from biomedical literature fed to a random forest model (Feature) on the validation set. The listed models were the best models encountered over 35 hyperparameter optimisation runs for each test and model type. We calculated the 95% confidence intervals (CIs) using bootstrap resampling with 1 000 bootstrap samples. A comparison between the test types was not possible, because the evaluated subsets differed significantly due to different user groups preferring to do certain tests in different amounts (Appendix 3.8.D).

	outbound		rest		return	
	CNN	Feature	CNN	Feature	CNN	Feature
AUC	**0.53** (0.50, 0.56)	0.50 (0.50, 0.53)	**0.53** (0.50, 0.56)	0.52 (0.50, 0.55)	**0.77** (0.74, 0.79)	**0.77** (0.75, 0.79)
AUPR	**0.60** (0.57, 0.64)	**0.60** (0.55, 0.62)	**0.62** (0.59, 0.66)	0.61 (0.55, 0.62)	0.72 (0.53, 0.87)	**0.86** (0.84, 0.88)

	voice		tapping		memory	
	CNN	Feature	CNN	Feature	RNN	Feature
AUC	0.53 (0.50, 0.55)	**0.56** (0.54, 0.58)	0.59 (0.57, 0.61)	**0.62** (0.60, 0.64)	**0.65** (0.60, 0.69)	0.52 (0.50, 0.57)
AUPR	**0.48** (0.45, 0.51)	0.45 (0.43, 0.48)	0.56 (0.53, 0.59)	**0.65** (0.62, 0.67)	**0.91** (0.88, 0.93)	0.87 (0.84, 0.89)

54

attention factors a_i using (Xu et al., 2015b; Schwab et al., 2017, 2019b):

$$a_i = \frac{\exp(u_i^T u_\mathrm{s})}{\sum_{j=1}^{m} \exp(u_j^T u_\mathrm{s})} \tag{3.3}$$

where

$$u_i = \text{activation}(W_\mathrm{s} h_i + b_\mathrm{s}) \tag{3.4}$$

Equation (3.4) corresponds to a single-layer MLP with a weight matrix W_s and bias b_s. The single-layer MLP projects h_i into a suitable hidden representation u_i for comparison with u_s. We then calculate the attention factors a_i by computing the softmax similarity of u_i to u_s. u_s is the most informative hidden representation, i.e. the hidden representation for which a_i would be the highest (Schwab et al., 2019b). W_s, b_s and u_s are learned parameters and jointly optimised with the other parameters during training. In equation 3.4, "activation" refers to an activation function. In the experiments in this work, we use the hyperbolic tangent function `tanh`.

3.4 Experiments

Our experiments aimed to answer the following questions:

(1) What is the comparative performance of various specialised models $P_\star$ in diagnosing PD from a single test?

(2) How do EAMs compare to existing methods for aggregating multiple local predictions?

(3) What is the overall diagnostic accuracy of our approach?

(4) Does the proposed hierarchical neural attention mechanism identify meaningful data points?

To answer these questions, we performed experimental comparisons between various baselines, predictive models and EAMs both on predicting PD from a single test and from an arbitrary number of tests.

Dataset and Study Cohort. We use data from the mPower study, a worldwide observational study about PD conducted entirely through smartphones (Bot et al., 2016). Starting in March 2015, the study recruited participants aged 18 and older around the world through a mobile app. Participants provided their demographic profile, including prior diagnoses of PD, through self-reporting, and performed the four test types regularly. Out of the study cohort, we used the subset of participants that were 45 or older, because there were very few participants in the dataset that had a clinical diagnosis at younger age. We used only those tests that were performed off medication, except for the memory tests. We performed a random split stratified by participant age to divide the available dataset into a training set (70%), validation set (10%), and test set (20%). Each participant and the tests they performed were assigned to exactly one of the three folds without any overlap (Table 3.2).

Models. For each test type, we trained several specialised predictive models $P_\star$ using both automated feature extraction with neural networks and random forest (RF) models. We used expert features from biomedical literature that have been shown to be predictive of PD in the given data modalities as inputs to the RF models. The complete list of features

Table 3.2: Population statistics of the training, validation, and test set. Numbers (#) shown are either absolute (Subjects), or the median with the associated 10% and 90% quantiles (in parentheses) over all subjects (Age, Walking, Voice, Tapping, Memory, Usage).

Property	Training	Validation	Test
Subjects (#)	1 314 (70%)	192 (10%)	347 (20%)
PD (%)	52.36	50.00	56.20
Female (%)	28.00	36.98	25.94
Age (years)	59.0 (47.00, 72.00)	60.00 (47.50, 70.50)	59.00 (46.50, 71.50)
Walking (#)	4.00 (1.00, 56.80)	5.00 (1.00, 120.80)	4.00 (1.00, 73.60)
Voice (#)	6.00 (1.00, 68.30)	6.50 (1.00, 115.70)	6.00 (1.00, 84.20)
Tapping (#)	6.00 (1.00, 58.70)	7.00 (2.00, 115.40)	7.00 (2.00, 81.00)
Memory (#)	3.00 (1.00, 38.20)	3.00 (1.00, 78.40)	3.00 (1.00, 55.00)
Usage (days)	5.58 (0.001, 82.17)	9.04 (0.002, 103.58)	6.04 (0.002, 102.44)

used for each test type can be found in Appendix 3.8.A. For the neural networks, we used different architectures of neural networks for each test depending on the type of input signal. For the walking, voice and tapping task, we used multi-layer convolutional neural networks (CNNs) with max pooling and temporal convolutions. For the memory test, we used a recurrent neural network (RNN) with bidirectional long short-term memory (BLSTM). Except for the voice test, the neural networks hosted the segment-level neural attention mechanisms described previously. For the voice CNN, we did not employ a neural attention mechanism because we found that it was detrimental to predictive performance. To implement the previously described EAM, we used a RNN architecture consisting of BLSTM cells. We trained EAMs using (1) only the RF models, (2) only the neural networks, and (3) an ensemble of both as specialised models to compare the performances of both approaches and whether their outputs are complementary. The detailed architectures for the neural networks and EAM are given in Appendix 3.8.B. The EAM received a one-hot encoded unique identifier of the specialised predictive model as input metadata $m_{\star,j}$ with each local per-test prediction $y_\star$. The unique identifier enabled the EAM to differentiate between the various specialised predictive models. We additionally tested passing timing information, including the time since the last performed test and the hour of day at which the test was performed, for each performed test. However, we found no performance benefit in adding timing information to the metadata. Lastly, in order to determine whether the use of an EAM improves performance over simpler approaches, we evaluated the performances of aggregating over local predictions $y_\star$ using the mean and maximum values of all local predictions. As a simple baseline based on demographics, we trained a MLP that received as input the age and gender of a participant and no information of any of their performed tests. To determine whether the proposed separation of learning to assess single test instances and learning to integrate multiple tests tests is beneficial, we also trained an end-to-end neural network jointly on both tasks. The end-to-end neural network used the same architectures as the specialised models to assess the tests and the same architecture as the EAM to integrate multiple tests.

Hyperparameters. We took a systematic approach to hyperparameter search. To avoid biases stemming from using different degrees of hyperparameter optimisation, we evaluated exactly 35 hyperparameter configurations for each trained specialised predictive model and EAM. We report the performances of those models which achieved the best validation set performances across the 35 runs. We selected the hyperparameters at random from a fixed range for each hyperparameter run. For the RF models, we used 512 to 1 024 trees in the forest and a maximum tree depth between 3 and 5. For all neural networks, we used dropout of 0 to 70% between hidden layers, an $L2$ penalty of 0, 0.0001 or 0.00001, and varying numbers of layers and hidden units depending on the test type (Appendix 3.8.C). For the EAM, we used 2 to 5 stacked BLSTM layers with 16 to 64 hidden units each. We optimised the neural networks' binary cross-entropy for up to 500 epochs with a learning rate of 0.0001, a batch size of 32, and an early stopping patience of 12 epochs on the validation set. For memory reasons, we used a batch size of 2 for the end-to-end trained neural network. All other hyperparameters and hyperparameter ranges were exactly the same as in the separated models.

Preprocessing. For computational performance reasons, we downsampled the input signals for the walking, voice and tapping test by factors of 10, 4, and 10, respectively. In our initial evaluation, we did not see significant differences in predictive performance when using higher resolution data. After downsampling, we zero-padded or truncated the size of the sensor data to fixed lengths for each test type if they were too short or too long, respectively. The fixed lengths were 300, 250, 25, and 300 samples per record for the walking, voice, memory and tapping tests, respectively. For the voice test, we did not pass the raw voice signal to the neural networks. Instead, we passed the Mel-frequency cepstral coefficients (MFCC) that we extracted from the audio signal using a window size of 30 ms, a stride of 10 ms and 40 coefficients as input signal. For the RFs, we used the raw audio signals downsampled from their original sampling rate of 44 100 Hz with factor 20 as inputs to the feature extractors. We standardised the accelerometer data

Figure 3.4: The outputs of the employed hierarchical neural attention mechanism on data from a user with PD that performed 18 tests. The timeline (top, left to right) shows all the tests performed by the user in temporal order. The tests performed (top, a_{test}) and the data segments within the tests (center, a_{seg}) were assigned attention weights that quantify their relative importance towards the final diagnostic score y. We show the outbound accelerometer data (left) and the rest accelerometer data (right) from walking test 10. In the outbound recording, the attention mechanism focused strongly on the long delay to start moving (segment 1), increasingly choppy movement while setting foot (segments 3, 4, and 5), and the abrupt stop (segment 7). In the rest recording, we found that the attention was evenly distributed across the recording, likely because the whole recording contained signs of what could have been resting tremor. Sightly more attention was paid to segments with increased motion (segments 5 and 13).

for the walking and tapping tests to have zero mean and unit variance.

Metrics. We computed the area under the receiver operating characteristic curve (AUC) and the area under the precision recall curve (AUPR) when comparing the different specialised predictive models. We evaluated the specialised models on the validation set to avoid test set leakage that could affect the evaluation of the models that aggregate information from multiple tests. We chose the best-performing specialised predictive models for each test for use in the aggregation models based on validation set performance. To compare the models that aggregated all available tests of a single user into a single diagnostic score, we additionally calculated the F1 score and the sensitivity at a fixed specificity level of 95%. Some of the data folds were not balanced between people with and without PD. In particular, comparing single-test performances between test types was not possible due to differences in the number of tests performed between people with and without PD (Appendix 3.8.D). We evaluated the performances of the three parts of the walking test (outbound, rest, and return) separately to determine their relative importances for diagnosing PD.

3.5 Results

Single-test Performance. In terms of single-test performance, we found that, generally, both RFs with expert features and automated feature extraction with neural networks achieved competitive results for all tests (Table 3.1). The performances of RFs with expert features and neural networks were similar across all tests, except for the tapping test, where the expert features significantly outperformed the neural networks, and the memory test, where the neural networks likewise outperformed the expert features. When comparing the three segments of the walking test, we found that return was the most informative for diagnosing PD.

Table 3.3: Comparison of the AUC, AUPR, F1, and sensitivity at a fixed specificity of 95% (Sens@95%Spec) on the test set of 347 participants across the methods that we evaluated. In parentheses are the 95% CIs calculated with 1 000 bootstrap samples.

Method	AUC	AUPR	F1	Sens@95%Spec
EAM (Both) + age + gender	**0.85** (0.81, 0.89)	**0.87** (0.82, 0.91)	0.81 (0.75, 0.85)	**0.43** (0.19, 0.54)
EAM (Neural networks) + age + gender	0.84 (0.80, 0.88)	0.86 (0.81, 0.90)	**0.82** (0.74, 0.86)	0.33 (0.21, 0.51)
EAM (Feature) + age + gender	0.84 (0.79, 0.88)	0.86 (0.81, 0.90)	0.76 (0.73, 0.84)	0.40 (0.23, 0.56)
End-to-end neural network + age + gender	0.50 (0.50, 0.56)	0.54 (0.46, 0.62)	0.27 (0.20, 0.70)	0.04 (0.01, 0.07)
Age + gender	0.74 (0.69, 0.79)	0.75 (0.68, 0.82)	0.72 (0.67, 0.79)	0.16 (0.09, 0.31)
EAM (Both)	0.70 (0.64, 0.75)	0.74 (0.66, 0.79)	0.67 (0.60, 0.71)	0.23 (0.15, 0.41)
EAM (Neural networks)	0.71 (0.65, 0.76)	0.75 (0.67, 0.80)	0.67 (0.61, 0.72)	0.24 (0.14, 0.41)
EAM (Feature)	0.71 (0.66, 0.76)	0.75 (0.67, 0.80)	0.68 (0.61, 0.73)	0.24 (0.14, 0.39)
Mean Aggregation (Neural networks)	0.64 (0.58, 0.69)	0.67 (0.58, 0.73)	0.60 (0.52, 0.68)	0.22 (0.10, 0.27)
Mean Aggregation (Feature)	0.62 (0.56, 0.68)	0.60 (0.51, 0.66)	0.62 (0.53, 0.69)	0.13 (0.00, 0.19)
Max Aggregation (Neural networks)	0.61 (0.55, 0.67)	0.61 (0.53, 0.68)	0.59 (0.54, 0.68)	0.03 (0.01, 0.19)
Max Aggregation (Feature)	0.61 (0.54, 0.66)	0.61 (0.52, 0.68)	0.60 (0.52, 0.65)	0.07 (0.03, 0.18)

Overall Performance. We found large differences in performance between the various aggregation models that took into account all the performed tests of a user (Table 3.3). Notably, EAMs outperformed all baselines by a large margin, and significantly improved upon the demographics-only model by integrating information from the tests performed by participants. We also found that expert features and neural network features were to some degree complementary, as the best EAM using both sets of predictive models outperformed its counterparts that only used one set of specialised predictive models. The neural networks trained end-to-end to simultaneously assess all types of tests and aggregate information from the available tests over time failed to converge. Closer analysis revealed that the end-to-end network was unable to effectively propagate gradients through the initially more attractive upper layers down to the per-test layers. Disentangling symptom assessment and temporal aggregation enabled EAMs to overcome this issue entirely.

Hierarchical Attention. We plotted the attributions of the hierarchical neural attention mechanism against the raw signals of a sample participant with PD (Figure 3.4). In the walking tests, the attributions potentially corresponded to regions where signs of resting tremor and rigid motions could have appeared. In the memory tests, we found that the focus was directed at the difficult end stage of the test (Appendix 3.8.E).

3.6 Discussion

Our work expands on prior studies (Arora et al., 2015) by developing an effective methodology for integrating evidence from multiple types of smartphone-based tests over long periods of time, introducing tools to identify the most salient data segments across the vast amounts of generated data points, and evaluating these novel approaches in a large, representative cohort. The availability of smartphone-based tools for diagnosing PD could have a profound impact on clinical practice by enabling clinicians to access long-term observational data on patients. These

additional data points could help give clinicians a more comprehensive and objective view on their patients' symptoms and symptom fluctuations, and therefore possibly enable more accurate diagnoses and treatment regimes. Another major potential benefit of enabling patients to record themselves with their smartphones is that it could enable clinicians to monitor their patients without requiring in-person visits that may be time-consuming and expensive, particularly in rural locations and developing countries. While our initial results are promising, further clinical validation is needed to determine whether the availability of smartphone data, the proposed diagnostic score, and in-depth information about the most relevant data points improve clinicians' ability to accurately diagnose PD.

Limitations. The main limitation of this work is that we use prior clinical diagnoses of users to train and evaluate our models. Clinical diagnoses for PD are themselves often inaccurate (Rizzo et al., 2016), and are therefore not a flawless gold standard to evaluate against. In addition, much like in clinical assessments, smartphone-based tests depend on PD symptoms being clearly distinguishable for at least some of the tests being performed. While smartphones enable patients to record themselves when they believe that their symptoms are most pronounced, they still might not be clearly distinguishable against normal human variability, particularly in early-onset PD. Furthermore, the accuracy of smartphone diagnostics may be reduced when confronted with other movement and neurologic disorders that may appear similar to PD. More data, ideally from a prospective study, is needed to conclusively determine the robustness of machine-learning and smartphone-based tests against these confounding factors.

3.7 Conclusion

We presented a machine-learning approach to distinguishing between people with and without PD from multiple smartphone-based tests. Our multistage approach is built on the idea of separately training (i) specialised models to assess symptom severity in instances of a single test, and

(ii) an EAM to integrate all available single-test assessments into a final diagnostic score. In addition, we introduced a hierarchical attention mechanism that shows both which tests out of all performed tests, and which segments within those tests were most important for the model's decision. We demonstrated experimentally that the presented approach leads to significant improvements over several strong baselines with an AUC of 0.85 (95% CI: 0.81, 0.89), an AUPR of 0.87 (95% CI: 0.82, 0.91) and a sensitivity at 95% specificity of 43% (95% CI: 0.19, 0.54) in data from a cohort of 1 853 participants. Our results confirm that machine-learning algorithms and smartphone data collected in the wild over extended periods of time could in the future potentially be used as a digital biomarker for the diagnosis of PD.

3.8 Supplementary Material

3.8.A Random Forest Features

The features used in the random forest (RF) models are listed in Tables 3.4, 3.5, 3.6, and 3.7. We chose our features based on prior research (Arora et al., 2015). We did not use all features reported in (Arora et al., 2015) because some of those features were too computationally inefficient to run in a reasonable amount of time in a dataset of the given size.

3.8.B Neural Network Architectures

Walking Test. For the walking test, we used a convolutional neural network (CNN) with temporal convolutions, a kernel size of 3, a number of sequential hidden layers, and an initial number of neurons with a growth rate per additional convolutional layer of 8. The number of initial neurons and the number of hidden layers were hyperparameters chosen at random from the ranges listed in Appendix 3.8.C. The convolutional layers were followed by an attention mechanism as described in the main body of the

paper, a single fully-connected layer with 32 neurons and an output neuron with a sigmoid activation. All layers except the output layer were followed by a batch normalisation layer and a leaky ReLU activation with a negative slope coefficient α of 0.3.

Voice Test. For the voice test, we used a CNN with spatial convolutions, a kernel size of 3, a number of sequential hidden layers, and an initial number of neurons with a growth rate per additional convolutional layer of 8. The number of initial neurons and the number of hidden layers were hyperparameters chosen at random from the ranges listed in Appendix 3.8.C. The convolutional layers were by three fully-connected layers with 512, 256 and 32 neurons, respectively, and an output neuron with a sigmoid activation. All layers except the output layer were followed by a batch normalisation layer and a leaky ReLU activation with a negative slope coefficient α of 0.3.

Tapping Test. For the tapping test, we used a recurrent neural network (RNN) for the tapping inputs with one bidirectional long short-term memory (BLSTM) layer, and a number of neurons for the BLSTM layer. The number of neurons for the BLSTM layer was a hyperparameter chosen at random from the range listed in Appendix 3.8.C. The recurrent layer was followed by an attention mechanism as described in the main body of the paper. All layers except the output layer were followed by a batch normalisation layer. In addition to the RNN for the tapping inputs, we used a CNN to jointly process the accelerometer signal. For the accelerometer neural network, we used a CNN with temporal convolutions, a kernel size of 3, a number of sequential hidden layers, and an initial number of neurons with a growth rate per additional convolutional layer of 4. The number of initial neurons and the number of hidden layers were hyperparameters chosen at random from the ranges listed in Appendix 3.8.C. The convolutional layers were followed by a single fully-connected layer with 32 neurons that was concatenated with the output of the RNN and then fed to an output neuron with a sigmoid activation.

Memory Test. For the memory test, we used a RNN with a number of BLSTM layers, and a number of neurons for the BLSTM layers. The number of BLSTM layers and the number of neurons for the BLSTM layers were hyperparameters chosen at random from the ranges listed in Appendix 3.8.C. The recurrent layers were followed by an attention mechanism as described in the main body of the paper. All layers except the output layer were followed by a batch normalisation layer.

Evidence Aggregation Model (EAM). For the EAM, we used a RNN with a number of BLSTM layers, and a number of neurons for the BLSTM layer. The number of BLSTM layers and the number of neurons for the BLSTM layers were hyperparameters chosen at random from the ranges listed in Appendix 3.8.C. The recurrent layers were followed by an attention mechanism as described in the main body of the paper, and an output neuron with a sigmoid activation. All layers except the output layer were followed by a batch normalisation layer.

3.8.C Hyperparameters

RF Hyperparameters. For the RF models, we varied the number of trees between 512 to 1 024 and the maximum depth of trees within the forest between 2 and 5.

Neural Network Hyperparameters. For the neural network models, we varied the dropout percentage after each hidden layer between 0 and 70%, and the $L2$ weight penalty from (0.0001, 0.00001, 0.0). For the EAMs, we varied the number of neurons per hidden layer between 16 to 64 and the number of hidden layers between 1 and 3. We also used different hyperparameter ranges for the specialised predictive models of each test type. For the walking test, we varied the number of neurons per hidden layer between 8 to 72 and the number of hidden layers between 5 and 7. For the voice test, we varied the number of neurons per hidden layer between 8 to 72 and the number of hidden layers between 5 and 6. For the tapping test,

we varied the number of neurons per hidden layer between 8 to 64 and the number of hidden layers between 6 and 8. For the memory test, we varied the number of neurons per hidden layer between 8 to 72 and the number of hidden layers between 1 and 2. We varied the numbers of hidden layers covered by the hyperparameter search depending on which type of neural network was used for the test (CNN or RNN), and depending on the total sequence length of the test's input signals.

3.8.D Per-test Population Statistics

The population statistics of the dataset folds used to train the specialised predictive models for each test type differed significantly (Tables 3.8, 3.9, 3.10, and 3.11). We split the subjects along the same folds as in the experiments for the aggregated models in order to prevent information leakage when training the specialised predictive models. This led to the evaluated folds for each test type being significantly different, as not all subjects in the per-test subsets performed at least one test of a given type, and, for some tests, different user groups (PD or control) preferred to do tests in different amounts. For example, the ratio of test set samples done by people with PD varies from 66% for the tapping test to 88% for the memory test. A direct comparison of the relative performances of the specialised predictive models between test types was therefore not possible, since the evaluation metrics are influenced not just by the performance differences between predictive models but also by the different underlying population statistics.

3.8.E Memory and Tapping Samples

We present typical samples of attention distributions for memory and tapping tasks in Figures 3.5 and 3.6, respectively.

Table 3.4: Features used as inputs to the random forests (RFs) used to assess walking tests. The features were calculated on both the x and z channel of the accelerometer data. The RF models did not use the gyroscope data.

	walking
Feature	Reference / Brief description
Mean	The mean of the amplitude of the input signal.
Standard deviation	The standard deviation of the amplitude of the input signal.
25% quartile	The 25% quartile of the amplitude of the input signal.
75% quartile	The 75% quartile of the amplitude of the input signal.
Inter-quartile range	The inter-quartile range of the amplitude of the input signal.
Median	The median of the amplitude of the input signal.
Range	The total range (max - min) of values of the input signal.
Skewness	The skewness of the amplitude of the input signal.
Kurtosis	The kurtosis of the amplitude of the input signal.
Mean squared energy	The mean squared energy of the amplitude of the input signal.
Entropy	The entropy of the input signal.
Mutual information	The mutual information of the input signal with the y-axis signal.
Detrended fluctuation analysis	(Arora et al., 2015)
Mean Teager-Kaiser energy operator	(Arora et al., 2015)
Cross-correlation	The cross-correlation of the input signal with itself up to lag level 1.
Zero-crossing rate	The zero-crossing rate of the input signal.

Table 3.5: Features used as inputs to the RFs used to assess voice tests. The features were calculated on the raw audio signal of the voice test. The RF models did not use the recordings taken during the countdown leading up to the voice test.

voice	
Feature	Reference / Brief description
Detrended fluctuation analysis	(Arora et al., 2015)
Mean Teager-Kaiser energy operator	(Arora et al., 2015)
Jitter	(Arora et al., 2015)
Shimmer	(Arora et al., 2015)
Pitch period entropy	(Arora et al., 2015)
Mel-frequency cepstral coefficients	(Arora et al., 2015)

Table 3.6: Features used as inputs to the RFs used to assess tapping tests. The features were calculated on the inter-tap intervals and tap positions.

tapping	
Feature	Reference / Brief description
Standard deviation	The standard deviation of the amplitude of the input signal.
Mean squared energy	The mean squared energy of the amplitude of the input signal.
Mean Teager-Kaiser energy operator	(Arora et al., 2015)
Cross-correlation	The cross-correlation of the input signal with itself up to lag level 2.
Detrended fluctuation analysis	(Arora et al., 2015)
$\text{Fatigue}_{10\%}$	(Arora et al., 2015)
$\text{Fatigue}_{25\%}$	(Arora et al., 2015)
$\text{Fatigue}_{50\%}$	(Arora et al., 2015)
Tremor between taps	(Arora et al., 2015)
Finger opening angle	(Arora et al., 2015)

70

Table 3.7: Features used as inputs to the RFs used to assess memory tests. The features were calculated on the inter-tap intervals and the button hit/miss time series. We additionally used the meta data associated with the memory test (overall score, number of games played, number of failures).

	memory
Feature	Reference / Brief description
Mean	The mean of the amplitude of the input signal.
Standard deviation	The standard deviation of the amplitude of the input signal.
Mean squared energy	The mean squared energy of the amplitude of the input signal.
Mean Teager-Kaiser energy operator	(Arora et al., 2015)
Cross-correlation	The cross-correlation of the input signal with itself up to lag level 2.
Detrended fluctuation analysis	(Arora et al., 2015)
Memory meta-data	The meta-data of the memory test.
$\text{Fatigue}_{10\%}$	(Arora et al., 2015)
$\text{Fatigue}_{25\%}$	(Arora et al., 2015)
$\text{Fatigue}_{50\%}$	(Arora et al., 2015)

Table 3.8: Population statistics of the training, validation, and test set for the walking tests. Numbers (#) shown are either absolute (Samples, Subjects), or the median with the associated 10% and 90% quantiles (in parentheses) over all subjects (Age, Usage).

Property	Training	Validation	Test
Samples (#)	8 461	1 496	2 119
PD (Samples, %)	57.94	59.09	68.19
Subjects (#)	609 (71%)	96 (11%)	151 (18%)
PD (Subjects, %)	43.19	44.79	47.68
Female (%)	26.77	34.38	19.21
Age (years)	59.00 (48.00, 72.00)	61.00 (49.00, 71.00)	58.00 (47.00, 71.00)
Usage (days)	9.06 (0.01, 95.98)	15.13 (0.01, 118.79)	6.04 (0.01, 114.08)

Table 3.9: Population statistics of the training, validation, and test set for the voice tests. Numbers (#) shown are either absolute (Samples, Subjects), or the median with the associated 10% and 90% quantiles (in parentheses) over all subjects (Age, Usage).

Property	Training	Validation	Test
Samples (#)	14 176	2 745	3 586
PD (Samples, %)	54.36	49.25	69.77
Subjects (#)	880 (70%)	141 (11%)	241 (19%)
PD (Subjects, %)	45.00	47.51	49.79
Female (%)	28.64	38.30	27.80
Age (years)	58.00 (47.00, 72.00)	60.00 (48.00, 70.00)	59.00 (46.00, 72.00)
Usage (days)	6.19 (0.003, 89.78)	14.15 (0.002, 102.97)	5.63 (0.003, 102.93)

Table 3.10: Population statistics of the training, validation, and test set for the tapping tests. Numbers (#) shown are either absolute (Samples, Subjects), or the median with the associated 10% and 90% quantiles (in parentheses) over all subjects (Age, Usage).

Property	Training	Validation	Test
Samples (#)	15 823	2 923	4 064
PD (Samples, %)	51.85	48.58	66.63
Subjects (#)	1 041 (70%)	158 (11%)	275 (19%)
PD (Subjects, %)	42.84	42.41	48.00
Female (%)	28.05	37.34	27.27
Age (years)	58.00 (47.00, 72.00)	59.00 (47.00, 70.00)	58.00 (46.00, 71.00)
Usage (days)	4.39 (0.001, 82.65)	7.28 (0.001, 88.80)	4.97 (0.001, 98.30)

Table 3.11: Population statistics of the training, validation, and test set for the memory tests. We included memory tests done on medication, because there were few tests done off medication. Numbers (#) shown are either absolute (Samples, Subjects), or the median with the associated 10% and 90% quantiles (in parentheses) over all subjects (Age, Usage).

Property	Training	Validation	Test
Samples (#)	4 720	1 143	1 600
PD (Samples, %)	88.62	86.18	87.81
Subjects (#)	337 (70%)	55 (11%)	91 (19%)
PD (Subjects, %)	64.09	65.45	75.82
Female (%)	35.01	27.27	28.57
Age (years)	61.00 (49.00, 73.00)	61.00 (51.00, 70.50)	63.00 (50.00, 72.00)
Usage (days)	44.57 (0.01, 152.89)	46.93 (1.20, 161.90)	38.84 (0.01, 162.26)

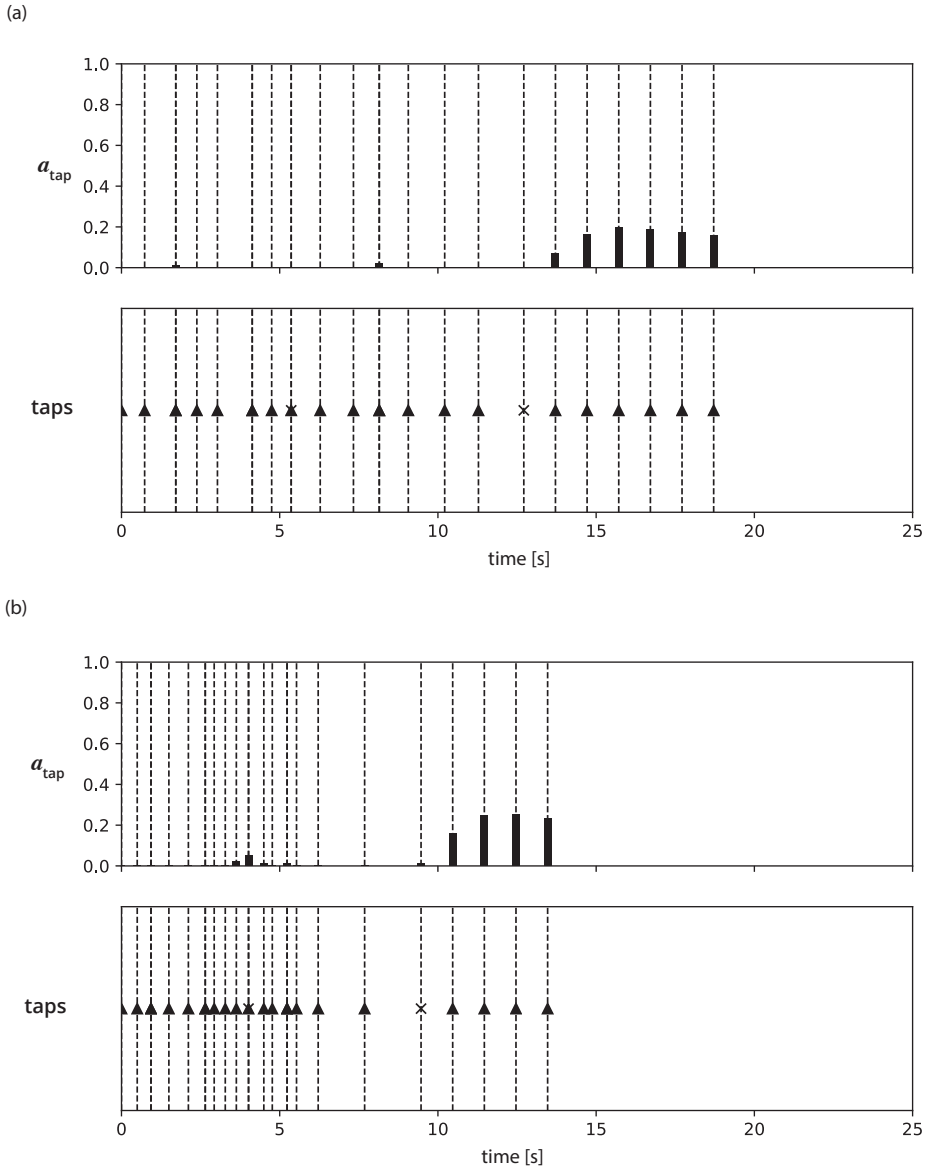


Figure 3.5: The outputs of the per-test neural attention mechanism (a_{tap}, top) on two representative samples (triangles and crosses, bottom) of memory tests from a user with Parkinson's disease. Triangles indicate correctly identified sequence elements and crosses indicate mistakes. We found that the predictive model's attention was typically focused on the more difficult final stage of the game. This pattern is visible in both samples (a) and (b). In both samples, we found that even mistakes in the early stage of the game do not receive a lot of attention relative to the more difficult end stage of the game.

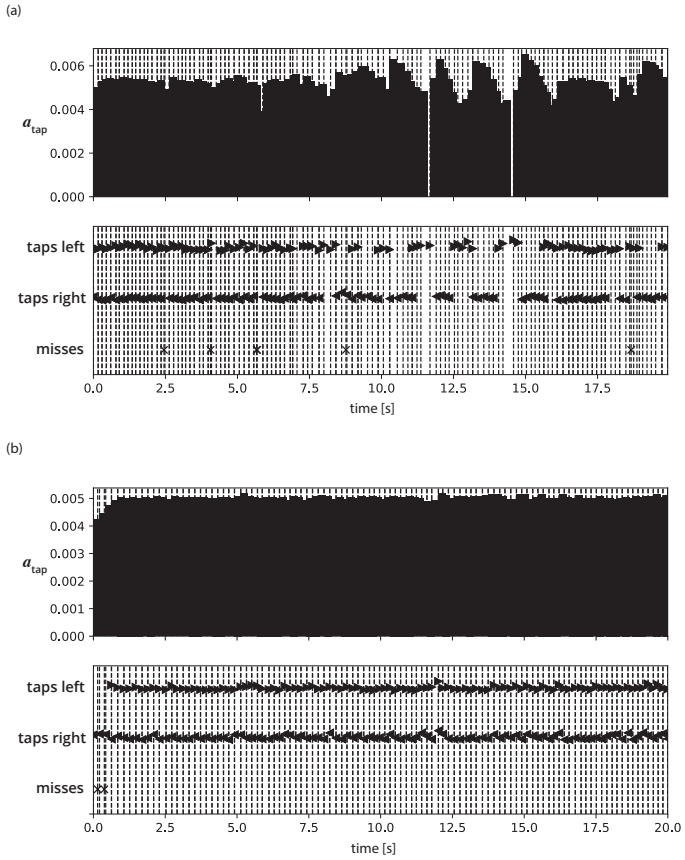


Figure 3.6: The outputs of the per-test neural attention mechanism (a_{tap}, top) on two representative samples (triangles and crosses, bottom) of tapping tests from a user with Parkinson's disease. The displacement of the tap triangles indicates the distance from the respective button's centre coordinate. In sample (a), we found that the assigned attention was typically lower in regions where the test was not performed properly (large breaks between taps, taps only on one side) and when taps were outside of the buttons' hit boxes (misses). In sample (b), we saw an almost uniform attention distribution, likely owing to the fact that the test was performed cleanly (aside from two misses at the start of the test). Our findings indicate that mistakes made in this test were not seen as predictive of Parkinson's disease. The predictive model instead focused on the user's overall tapping performance when the test was performed as intended. Furthermore, the predictive model distributed its attention among large numbers of taps, indicating that features spanning over many taps were in general seen as more predictive than individual taps.

This page was intentionally left blank.

Chapter 4

Learning Important Features with Neural Networks

"A wise man proportions his belief to the evidence." - David Hume

Knowledge of the importance of input features towards decisions made by machine-learning models is essential to increase our understanding of both the models and the underlying data. Here, we present a new approach to estimating feature importance with neural networks based on the idea of distributing the features of interest among experts in an attentive mixture of experts (AME). AMEs use attentive gating networks trained with a Granger-causal objective to learn to jointly produce accurate predictions as well as estimates of feature importance in a single model. Our experiments show (i) that the feature importance estimates provided by AMEs compare favourably to those provided by state-of-the-art methods, (ii) that AMEs are significantly faster at estimating feature importance than existing methods, and (iii) that the associations discovered by AMEs are consistent with those reported by domain experts.

4.1 Introduction

Neural networks are often criticised for being black-box models (Castelvecchi, 2016). Researchers have addressed this criticism by developing tools that provide visualisations and explanations for the decisions of neural networks (Baehrens et al., 2010; Simonyan et al., 2014; Zeiler & Fergus, 2014; Xu et al., 2015b; Shrikumar et al., 2017; Krause et al., 2016; Montavon et al.,

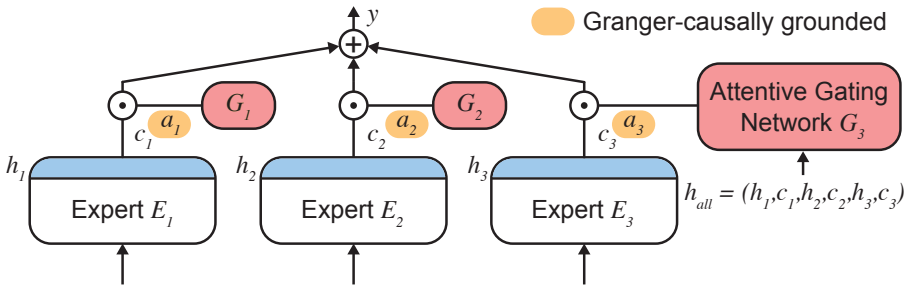


Figure 4.1: An overview of attentive mixtures of experts (AMEs). The attentive gating networks G_i (red) attend to the combined hidden state h_{all} (blue) of the AME. Each expert's G_i assigns an attentive factor a_i to opportunistically control its contribution c_i to the AME's final prediction y.

2017; Koh & Liang, 2017). These explanations are desirable for the many machine-learning use cases in which both predictive performance and interpretability are of paramount importance (Kindermans et al., 2017; Smilkov et al., 2017; Doshi-Velez & Kim, 2017). They enable us to argue for the decisions of machine-learning models, show when algorithmic decisions might be biased or discriminating (Hardt et al., 2016), help uncover the basis of decisions when there are legal or ethical circumstances that call for explanations (Goodman & Flaxman, 2017), and may facilitate the discovery of patterns that could advance our understanding of the underlying phenomena (Shrikumar et al., 2017).

Estimating the relative contribution of individual input features towards outputs of a deep neural network is hard because the input features undergo multiple hierarchical, interdependent and non-linear transformations as they pass through the network (Montavon et al., 2017). We propose a new approach to feature importance estimation that optimises jointly for predictive performance and accurate assignment of feature importance in a single end-to-end trained neural network. Structurally, our approach builds on the idea of distributing the features of interest among experts in a mixture of experts (Jordan & Jacobs, 1994). The mixture of experts uses attentive gating networks to assign importance weights to individual experts (Figure 4.1). However, when trained naïvely, this structure alone does not generally learn to accurately assign weights that correspond to the importance of the

Table 4.1: Comparison of AMEs to several representative methods for feature importance estimation.

	AME	Attention	DeepLIFT	LIME	SHAP
Model-agnostic	×	×	×	✓	✓
Measure of expected quality	✓	×	×	×	×
Computation time (AME = 1x)	1x	1x	2x	100-1000x	>1000x

experts' input features. We therefore draw upon a previously unreported connection between Granger-causality and feature importance estimation to define a secondary Granger-causal objective. Using the Granger-causal objective, we ensure that the weights given to individual experts correlate strongly and measurably with their ability to contribute to the decision at hand. Our experiments demonstrate that this optimisation-based approach towards learning to estimate feature importance leads to improvements of several orders of magnitude in computational performance over state-of-the-art methods. In addition, we show that the Granger-causal objective correlates with the expected quality of importance estimates, that AMEs compare favourably to the best existing methods in terms of feature importance estimation accuracy, and that AMEs discover associations that are consistent with those reported by domain experts.

Contributions. This chapter contains the following contributions:

- We delineate an end-to-end trained AME that uses attentive gating to assign weights to individual experts.
- We introduce a Granger-causal objective that measures the degree to which assigned feature importances correlate with the predictive value of features towards individual decisions.
- We compare AMEs to state-of-the-art importance estimation methods on three datasets. The experiments show that AMEs are significantly faster than existing methods, that AMEs compare favourably to existing methods in terms of attribution accuracy, and that the associations discovered by AMEs are consistent with human experts.

4.2 Related Work

There are four main categories of approaches to assessing feature importance in neural networks:

Perturbation-based Approaches. Perturbation-based approaches attempt to explain the sensitivity of a machine-learning model to changes in its inputs by modelling the impact of local perturbations (Ribeiro et al., 2016b; Adler et al., 2016; Fong & Vedaldi, 2017; Lundberg & Lee, 2017). Examples of perturbation-based approaches are local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016b) and Shapley additive explanations (SHAP) (Lundberg & Lee, 2017). The main drawbacks of perturbation-based approaches are (a) that perturbed samples might not be part of the original data distribution (Ribeiro et al., 2018), and (b) that they are computationally inefficient, as hundreds to thousands of model evaluations are required to sample the space of local perturbations. Perturbation-based approaches are applicable to any machine-learning model (Ribeiro et al., 2016b; Lundberg & Lee, 2017).

Gradient-based Methods. Gradient-based approaches are built on the idea of following the gradient from the output nodes of a neural network to the input nodes to obtain the features that the output was most sensitive to (Baehrens et al., 2010; Simonyan et al., 2014). Gradient-based approaches are therefore only applicable to differentiable models. Several improvements to this technique have since been proposed (Zeiler & Fergus, 2014; Smilkov et al., 2017; Sundararajan et al., 2017). In particular, (Shrikumar et al., 2017) introduced the deep learning important features (DeepLIFT) method that addresses the issue of saturating gradients.

Attentive Models. Attentive models have been used in various domains to improve both interpretability (Xu et al., 2015b; Choi et al., 2016b; Schwab et al., 2017; Schwab & Karlen, 2019b) and performance (Bahdanau et al.,

2015; Yang et al., 2016). In computer vision, related works used attention in convolutional neural networks (CNNs) that selectively focus on input data (Ba et al., 2014) and internal convolutional filters (Stollenga et al., 2014). However, fundamentally, naïve soft attention mechanisms do not provide any incentive for a neural network to yield attention factors that correlate with feature importance. When used on top of recurrent neural networks (RNNs), attention mechanisms may propagate information across time steps through the recurrent state, and are therefore not guaranteed to accurately represent importance (Sundararajan et al., 2017).

Mimic Models. Mimic models are interpretable models trained to match the output of a black-box model. Rule-based models (Andrews et al., 1995) and tree models (Schwab & Hlavacs, 2015; Che et al., 2016) are examples of models that have been used as mimic models. Mimic models are not guaranteed to match the original model and may thus not be truthful to the original model.

Model-agnosticism. Irrespective of the category of the approach, we would ideally want a feature importance estimation method to be model-agnostic, i.e. independent of the choice of predictive model (Ribeiro et al., 2016a). The main arguments for model-agnosticism are flexibility to choose the predictive model and feature representation as necessary (Ribeiro et al., 2016a). However, in practice, the generality of model-agnostic approaches comes at a considerable cost in computational performance and scalability (Table 4.1). With datasets continuously growing in size and neural networks becoming the preferred choice of model in many domains, model-specific feature importance estimation methods are often the only viable choice. This is evidenced by the recent surge in works applying model-specific approaches to analyse predictive relationships in large-scale datasets (Esteva et al., 2017; Ilse et al., 2018). In addition, it would be desirable to have a measure of the expected quality of the provided importance estimates. Against the backdrop of estimates that are potentially not truthful to the underlying data, such a measure would enable us to assess

the expected estimation accuracy and inform us when accurate estimates can not be expected. To the best of our knowledge, the presented Granger-causal objective is both the first tool that quantifies the expected quality of importance estimates, and the first objective that enables neural networks to learn to estimate feature importance.

4.3 Attentive Mixtures of Experts

We consider the setting in which we are given a dataset containing training samples X. Each X consists of input features x_i with $i \in [1 \; .. \; p]$ where p is the number of input features per sample. A ground truth label y_{true} is available for each training sample. Using the labelled training dataset, we wish to train a model that produces (1) accurate output predictions y for new samples for which we do not have labels, and (2) feature importance scores a_i that correspond to the importance of each respective input feature x_i towards the output y. To model this problem setting, we introduce AMEs, a mixture of experts model that consist of p experts E_i and their corresponding attentive gating networks G_i (Figure 4.1). At prediction time, the attentive gating networks output an attention factor a_i for each expert to control its respective contribution c_i to the AME's final prediction y. All of the experts and the attentive gating networks are neural networks with their own parameters and architectures[1]. AMEs do not impose any restrictions on the experts other than that they need to expose their topmost feature representation h_i and their contribution c_i for a given X.

As input to the gating networks, the hidden states h_i and local contributions c_i of each expert are concatenated to form the combined hidden state h_{all} of the whole AME:

$$h_{\text{all}} = \text{concatenate}(h_1, c_1, h_2, c_2, ..., h_p, c_p) \tag{4.1}$$

[1]The experts are, however, not separate models because all parts of AMEs are connected, differentiable, and trained end-to-end. An AME is therefore a single model and not an ensemble of models.

We denote c_i as the output $E_i(x_i)$ of the ith expert for the given feature of the input data x_i. a_i represents the output $G_i(h_{\text{all}})$ of the ith attentive gating network G_i with respect to the combined hidden state h_{all} of the AME. The output y of an AME is then given by:

$$y = \sum_{i=1}^{p} \underbrace{G_i(h_{\text{all}})}_{a_i} \underbrace{E_i(x_i)}_{c_i} \tag{4.2}$$

The attention factors a_i modulate the contribution c_i of each expert to the final prediction y based on the AME's combined hidden state h_{all}. The attention factors therefore represent the importance of each expert's contribution towards the output y. The motivation behind structuring AMEs as a mixture of experts with input features x_i distributed across experts is to ensure (1) that each expert's contribution c_i can *only* be based on their respective input feature x_i, and (2) that the importance of c_i towards the final prediction y can only be increased by increasing its respective attention factor. Splitting the features across experts guarantees that there can not be any information leakage across features, and that the attention factors a_i can in turn safely be interpreted as the importance of the input feature x_i towards y upon model convergence. We calculate the attention factors a_i using:

$$a_i = \frac{\exp(u_i^T u_{s,i})}{\sum_{j=1}^{p} \exp(u_j^T u_{s,i})} \tag{4.3}$$

where

$$u_i = \text{activation}(W_i h_{\text{all}} + b_i) \tag{4.4}$$

corresponds to a single-hidden-layer multi-layer perceptron (MLP) with an activation function, a weight matrix W_i and bias b_i. To compute the attention factors, we first feed the combined state h_{all} of the AME into the MLP to get u_i as a projected hidden representation of h_{all} (Xu et al., 2015b; Rocktäschel et al., 2016; Yang et al., 2016). We then compute the similarity of the projected hidden representation u_i to a per-expert context vector $u_{s,i}$. The context vector $u_{s,i}$ can be seen as a fixed high-level representation that answers

the question: "What projected hidden representation would be the most informative for this expert?". The per-expert context vector $u_{s,i}$ is initialised randomly and has to be learned with the other network parameters during training. We obtain normalised importance scores a_i from the similarities through a softmax function (Eq. 4.3). The attention factors a_i are used to weight the contributions c_i of each expert towards the final decision y of the AME (Eq. 4.2). The soft attention mechanism formulated in Equations 4.3 and 4.4 closely follows the definitions used in related works (Xu et al., 2015b; Rocktäschel et al., 2016; Yang et al., 2016) with two notable exceptions: Firstly, we use h_{all} rather than just the hidden representation of a single expert as input to the soft attention mechanism. This enables the AME to simultaneously take into account the information from all available experts when producing its attention factors a_i and its prediction y, despite not sharing features across the experts themselves. Secondly, we use a separate attentive gating network for each expert to produce the attention factors. This is in contrast to existing works that use a shared representation either over feature maps in a CNN for image data (Xu et al., 2015b) or over the hidden states of a RNN for sequence data (Xu et al., 2015b; Choi et al., 2016b; Yang et al., 2016). Using a shared or overlapping attention mechanism is problematic for importance estimation, as information from features could potentially leak across features. This is best exemplified by attention mechanisms on top of RNNs, where information can propagate across time steps through the recurrent state, and therefore influence the model output through means other than the attention factor (Sundararajan et al., 2017). The use of separate attention mechanisms prevents information leakage entirely, and at the same time enables us to apply soft attention to non-sequential and non-spatial input data.

4.4 Granger-causal Objective

A fundamental issue of naïvely-trained soft attention mechanisms is that they provide no incentive to learn feature representations that yield accurate attributions (Sundararajan et al., 2017). Naïvely-trained attentive gating

networks may therefore not accurately represent feature importance or even collapse towards attractive minima, such as assigning an attention weight of $a_i = 1$ to a single expert and 0 to all others (Bengio et al., 2015; Shazeer et al., 2017). To ensure the assigned attention weights correspond to feature importance, we introduce a secondary objective function that measures the mean Granger-causal error (MGE). Granger-causality follows the Humean definition of causality that, under certain assumptions, declares a causal relationship $X \rightarrow Y$ between random variables X and Y if we are better able to predict Y using all available information than if the information apart from X had been used (Granger, 1969). Given input sample X, we denote $\varepsilon_{X\setminus\{i\}}$ as the AME's prediction error without including any information from the ith expert and ε_X as the AME's prediction error when considering all available information. To estimate $\varepsilon_{X\setminus\{i\}}$ and ε_X, we use differentiable auxiliary predictors $\mathrm{P}_{\text{aux},i}$ and $\mathrm{P}_{\text{aux},c}$ that receive as input the concatenated hidden representations of all experts excluding the ith expert's hidden representation $h_{\text{all}} \setminus h_i$ and the concatenated hidden representations of all experts h_{all}, respectively. The auxiliary predictors are trained jointly with the AME.

$$y_{\text{aux},i} = \mathrm{P}_{\text{aux},i}(h_{\text{all}} \setminus h_i) \tag{4.5}$$

$$y_{\text{aux},c} = \mathrm{P}_{\text{aux},c}(h_{\text{all}}) \tag{4.6}$$

We then calculate $\varepsilon_{X\setminus\{i\}}$ and ε_X by comparing the auxiliary predictions $y_{\text{aux},i}$ and $y_{\text{aux},c}$ with the ground truth labels y_{true} using the auxiliary loss function $\mathcal{L}_{\text{aux}}$. We use the mean absolute error as $\mathcal{L}_{\text{aux}}$ for regression problems and categorical cross-entropy for classification problems.

$$\varepsilon_{X\setminus\{i\}} = \mathcal{L}_{\text{aux}}(y_{\text{true}}, y_{\text{aux},i}) \tag{4.7}$$

$$\varepsilon_X = \mathcal{L}_{\text{aux}}(y_{\text{true}}, y_{\text{aux},c}) \tag{4.8}$$

Following (Granger, 1969), we define the degree $\Delta\varepsilon_i$ to which the ith expert contributes to the output y as the decrease in error associated with adding

that expert's information to the set of available information sources:

$$\Delta\varepsilon_{X,i} = \varepsilon_{X\setminus\{i\}} - \varepsilon_X \tag{4.9}$$

This definition of $\Delta\varepsilon_{X,i}$ naturally resolves cases where combinations of features enable improvements in the prediction error - both experts would be attributed equally for the decrease. We then normalise the desired attribution ω_i corresponding to the ith experts' attention weights a_i.

$$\omega_i(X) = \frac{\Delta\varepsilon_{X,i}}{\sum_{j=1}^{p}\Delta\varepsilon_{X,j}} \tag{4.10}$$

Where equation (4.10) normalises the attributions across all experts to ensure that they are on the same scale across decisions. We calculate the Granger-causal objective $\mathcal{L}_{\text{MGE}}$ by computing the average probabilistic distance over n samples between the target distribution Ω, with $\Omega(i) = \omega_i$, and the actual distribution A, with $A(i) = a_i$, of attention values using a distance measure D. The Kullback-Leibler divergence (Kullback, 1997) is a suitable differentiable D for attention distributions (Itti & Baldi, 2006).

$$\mathcal{L}_{\text{MGE}} = \frac{1}{n}\sum_{X}\text{D}(\Omega, A) \tag{4.11}$$

Because the Granger-causal loss measures the average probabilistic distance of the actual attributions to the desired Granger-causal attributions, it is valid to use it as a proxy for the expected quality of explanations. A Granger-causal loss of 0 indicates a perfect match with the Granger-causal attributions. We can therefore apply the familiar framework of cross-validation and held-out test data to estimate the expected quality of the importance estimates a_i on unseen data. Finally, the total loss $\mathcal{L}$ is the sum of the main loss and the Granger-causal loss weighted by a hyperparameter α. We chose linear mixing of the loss terms as a simple starting point. We note that more complex methods for combining the two loss terms are potentially interesting avenues for future research.

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{\text{main}} + \alpha\mathcal{L}_{\text{MGE}} \tag{4.12}$$

The core idea of the Granger-causal objective is to train predictors on distinct subsets of the input data to measure how much the exclusion of individual features reduces model performance. This approach to importance estimation is not new (Štrumbelj et al., 2009) and is commonly practiced in ablation studies. In addition, a similar approach, called Shapley value analysis (Lipovetsky & Conklin, 2001) or Shapley regression values (Lundberg & Lee, 2017), has been proposed for regression using the game-theoretic concept of Shapley values (Shapley, 1953; Lundberg & Lee, 2017). The main difference between Shapley values and Granger-causality is that feature importance in Shapley values is defined as the marginal contribution towards the model output whereas Granger-causality defines importance in terms of the marginal contribution towards the reduction in prediction error. This subtle change in definition improves computational and memory scalability from factorial to linear in the number of features as we only have to train one additional auxiliary model per feature rather than one for every possible subset of features (Lipovetsky & Conklin, 2001; Lundberg & Lee, 2017).

4.5 Experiments

To compare AMEs to state-of-the-art methods for importance estimation, we performed experiments on an established benchmark for importance estimation and two real-world tasks. Our goal was to answer the following questions:

(1) How do AMEs compare to state-of-the-art feature importance estimation methods for neural networks in terms of estimation accuracy and computational performance?

(2) Does jointly optimising AMEs for predictive performance and accurate estimation of feature importance have an adverse impact on predictive performance?

(3) Does a lower test-set MGE correlate with a better expected estimation

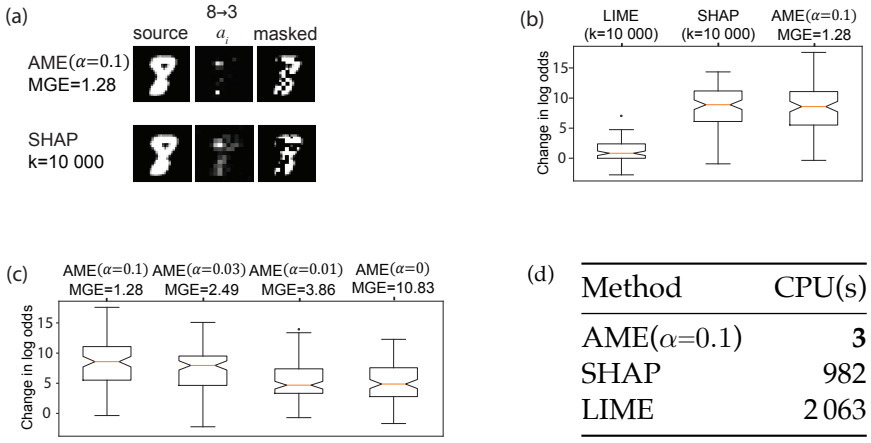
accuracy on unseen data?

(4) How do varying choices of α impact predictive performance and attribution accuracy?

(5) Are the associations identified by AMEs and other methods consistent with those reported by domain experts?

4.5.1 Important Features in Handwritten Digits

We performed the MNIST (LeCun et al., 1998) benchmark proposed by (Shrikumar et al., 2017) to compare AMEs to LIME (Ribeiro et al., 2016b) and SHAP (Lundberg & Lee, 2017), and to validate whether a lower test-set MGE on test data indicates a better estimation accuracy. Because AMEs provide a single set of importance scores per decision and not one set of importance scores for each possible output class, we adapted the benchmark to use a binary classifier that was trained to distinguish between a source and a target digit class ($8 \rightarrow 3$). We used LIME, SHAP and multiple AMEs to determine the most important pixels in an image of the source digit. The most important pixels in this setting corresponded to those pixels which distinguish the source digit from the target digit. We masked the top 10% of most important pixels (Figure 4.2a) and calculated the change in log odds for classifying across $n = 100$ samples (Figure 4.2b) to quantify to what degree the feature importance estimation methods were able to identify the important pixels for distinguishing the two digit classes (Shrikumar et al., 2017; Lundberg & Lee, 2017). We brought LIME and SHAP to the same scale as the AME's attention factors a_i by applying the normalising transform $a_i = \frac{|e_i|}{\sum_{j=0}^{n}|e_j|}$ (eq. 4.13). We trained AME(α=0) and AME(α=0.1) until convergence (100 epochs, 6 epochs early stopping patience) and stopped the training of AME(α=0.03) and AME(α=0.01) prematurely after 10 epochs to obtain AMEs with higher test-set MGE values for comparison (Figure 4.2c). We applied LIME and SHAP to the AME(α=0.1) with $k = 10000$ samples. Appendix 4.7.C lists architectures and hyperparameters.



(d)

Method	CPU(s)
AME(α=0.1)	**3**
SHAP	982
LIME	2 063

Figure 4.2: Determining important features on MNIST. (a) The attention map a_i shows which pixels were assigned the most importance. We masked the most important pixels to change the prediction to the target digit (more samples in Appendix 4.7.D). AMEs were (b) comparable to SHAP in the change in log odds (d) at significantly lower runtime when masking over $n = 100$ random images. (c) Lower MGEs correlated with better estimates when comparing AMEs with different levels of test set MGE.

4.5.2 Drivers of Medical Prescription Demand

To gain a deeper understanding of what factors drive prescription demand, we trained machine-learning models to predict the next month's demand for prescription items.

Dataset. We used data related to prescription demand in England, United Kingdom during the time frame from January 2011 to December 2012. We used data streams split into six feature groups: (a) demand history, (b) online search interest, (c) regional weather, (d) regional demographics, (e) economic factors and (f) labor market data. Appendix 4.7.E contains a description of the dataset and the list of input features per expert (total number of features $p = 585$). We applied a random split by practice to separate the data into training (60%, 5 673 practices, 24.30 million time series), validation (20%, 1 891 practices, 9.07 million time series) and test set (20%, 1 891 practices, 9.07 million time series). Because LIME and SHAP did

not scale to the size of this test set, we used a subset of 3 practices (17 316 time series) to perform the comparison on importance estimation speed.

Models. We trained autoregressive integrated moving average (ARIMA) models, recurrent neural networks (RNN), feedforward neural networks (FNN), and AMEs trained with (α>0) and without (α=0) the Granger-causal objective. Each feature group was represented as an expert in the AMEs for a total of six expert networks. The AMEs trained without the Granger-causal objective served as a baseline of relying on neural attention only. ARIMA served as a baseline that did not make use of any information apart from the revenue history. We applied all feature importance estimation methods except DeepLIFT to the same AME(α=0.5). Because there, to our knowledge, currently exists no DeepLIFT propagation rule for attentive gating networks, we used the highest-performing FNN to produce the DeepLIFT explanations (architectures in Appendix 4.7.F).

Hyperparameters. For all neural networks, we performed a hyperparameter search with hyperparameters chosen at random from predefined ranges ($L = 1-3$ hidden layers, $N = 8-128$ hidden units per layer, $0-80\%$ dropout) over 35 runs. We selected those models from the hyperparameter search that achieved the best performance. Methodologically, we optimised the neural networks' mean squared error (MSE), batch size of 256, with an early stopping patience of 12 epochs and a learning rate of 0.0001. For ARIMA, we

Table 4.2: Comparison of the symmetric mean absolute percentage error (SMAPE; in $\%$) on the test set of 1 891 practices ($n = 9.07$ million time series), and the average $\pm$ standard deviation of CPU hours used for training and evaluation across the 35 runs.

Method	SMAPE (%)	CPU (hr)
RNN	**32.79**	0.25 ± 0.07
FNN	32.87	$\mathbf{0.06 \pm 0.02}$
AME(α=0)	33.08	0.45 ± 0.14
AME(α=0.04)	33.85	0.21 ± 0.08
ARIMA	34.98	527.96

used the iterative parameter selection algorithm from (Hyndman & Khandakar, 2007). To better understand the impact of α, we trained AMEs with $\alpha \in [0, 0.1]$ chosen on a grid in steps of 0.01. We used a neighbourhood of $k = 2000$ perturbed samples for LIME. Despite our use of a small subset for the comparison on estimation speed, we were only able to apply SHAP with $k = 5$ perturbed samples. The expected computation time for applying SHAP with $k = 100$ perturbed samples was 9 months of CPU time.

Pre- and Postprocessing. Prior to fitting the models, we normalised the prescription revenue history data for each time series to the range $[0, 1]$. We standardised all other features to have zero mean and unit variance.

Metrics. We compared the predictive accuracy of the different models by computing their symmetric mean absolute percentage error (SMAPE) (Flores, 1986) on the test set of 1 891 practices. We additionally compared the speed of the various feature importance estimation methods by measuring the computation time in CPU seconds used for evaluation and the time in CPU hours used for training.

Results. For importance estimation, AMEs (2 CPU seconds) were faster than DeepLIFT (24 CPU seconds), LIME (10 464 CPU seconds) and SHAP (729 068 CPU seconds) by one, four and six orders of magnitude, respectively. In terms of predictive performance, AME(α=0) models performed slightly worse than the FNN and RNN (Table 4.2). Furthermore, AME(α=0.04) performed worse than AME(α=0). This indicates that there was a small performance decrease associated with both (i) the use of attentive gating networks, and (ii) optimising jointly to maximise predictive performance and feature importance estimation accuracy. We hypothesise that (ii) is caused by adverse gradient interactions (Doersch & Zisserman, 2017; Schwab et al., 2018a) between the main task and the Granger-causal objective. We also found that AMEs indeed effectively learn to match the desired Granger-causal attributions (Eq. 4.10) with a Pearson correlation r^2 of 0.84 measured on the test set. In contrast, the AME(α=0) trained without

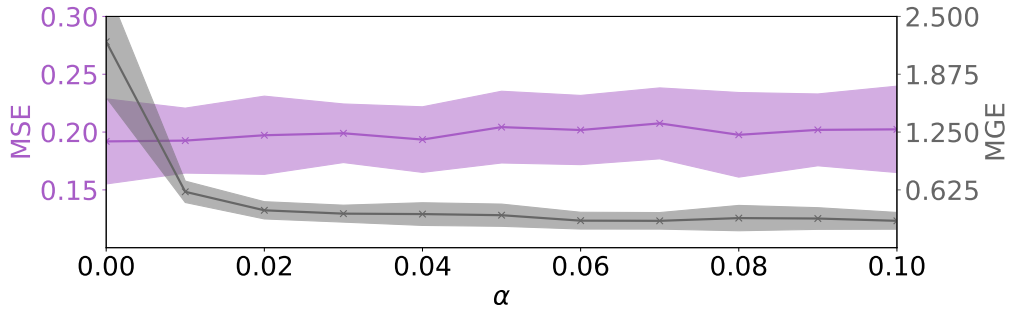


Figure 4.3: The mean value (solid lines) and the standard deviation (shaded area) of the MSE (purple) and the MGE (grey) of AMEs trained with varying choices of $\alpha \in [0, 0.1]$ across 35 runs on the test set of 1 891 practices.

the Granger-causal objective only reached a r^2 of 0.19. The training time of AME($\alpha = 0.04$) was comparable to RNNs.

Impact of α. Increasing values of α in the range of $[0, 0.1]$ lead to an exponential improvement in MGE that was accompanied with a minor decrease in MSE (Figure 4.3). A good middle ground was at $\alpha \approx 0.03$, where roughly 80% of the attribution accuracy gains were realised while maintaining most of the performance. The relationship between MGE and MSE was constant for values of $\alpha > 0.1$.

4.5.3 Discriminatory Genes Across Cancer Types

To pinpoint the genes that differentiate between several types of cancer, we analysed the feature importances in machine-learning models trained to classify gene expression data as being either breast carcinoma (BRCA), kidney renal clear cell carcinoma (KIRC), colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD) and prostate adenocarcinoma (PRAD).

Dataset. We used gene expression data from multiple cancer types in 801 individuals from The Cancer Genomic Atlas (TCGA) RNAseq dataset. To

Table 4.3: Comparison of the number of gene-cancer associations that were substantiated by literature evidence in the top 10 genes by average importance (Recall@10), and the number of CPU seconds used to compute them.

Method	Recall@10	CPU(s)
AME(α=0.05)	**10**	**3**
RF	**10**	12
SHAP(k=100)	8	6119
LIME(k=400)	8	80
DeepLIFT	7	6
AME(α=0)	2	**3**

keep visualisations succinct, we used a subset of 100 genes as input data. We applied a stratified random split to separate the data into training (60%, 480 samples), validation (20%, 160 samples) and test set (20%, 161 samples).

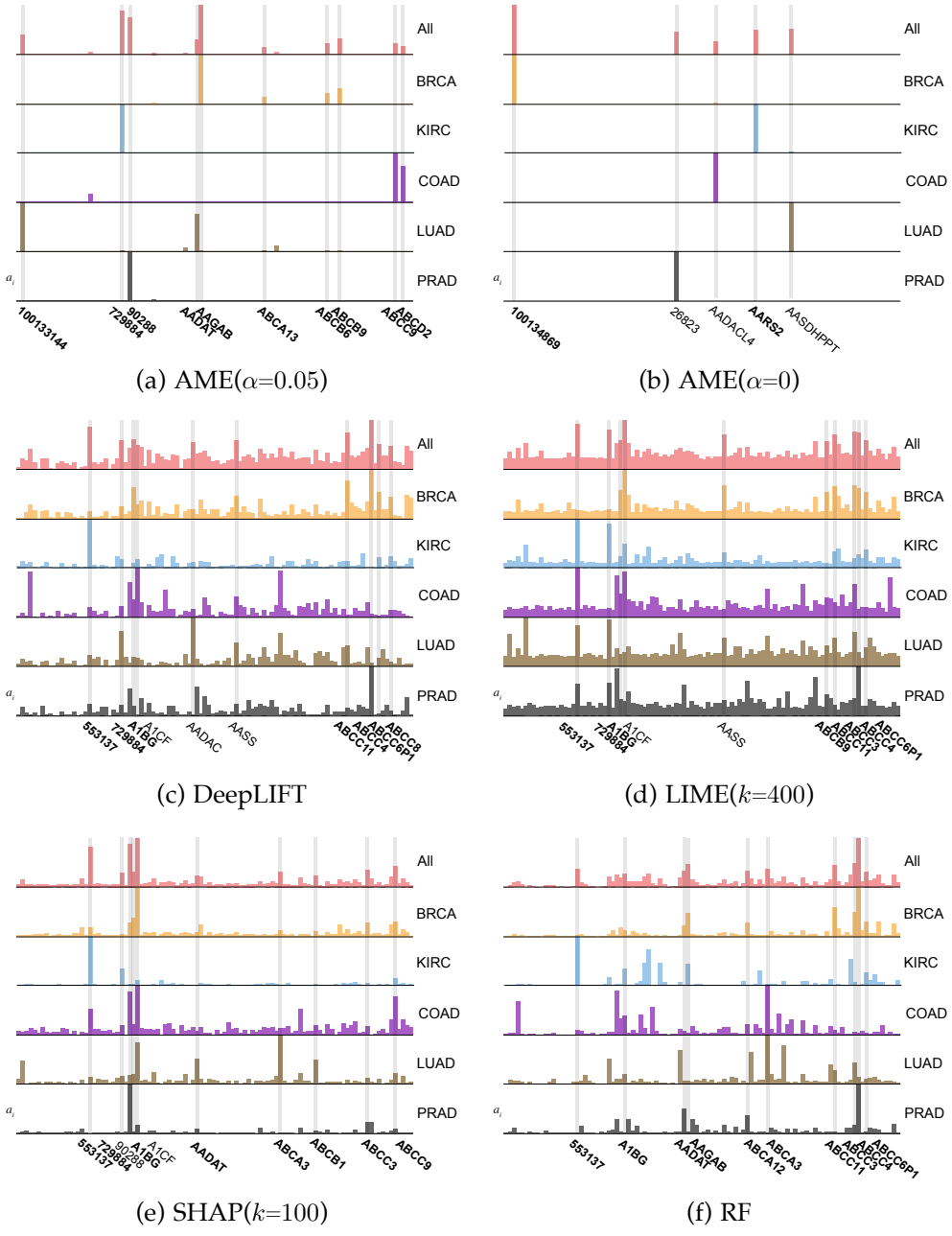
Models. We trained FNN, AME(α=0) and AME(α=0.05) (architectures in Appendix 4.7.G). LIME(k=400) and SHAP(k=100) were applied to the AME(α=0.05) and DeepLIFT to the best FNN for the same reason as in experiment 2. We also trained five random forests (RF) (Breiman, 2001) with 2 048 trees in a binary one-vs.-all classification setting for each cancer type. We used the Gini importance measure (Breiman, 2001; Genuer et al., 2010; Louppe et al., 2013) derived from the RFs as a baseline that was independent of the neural networks.

Hyperparameters. For each of the 100 gene loci, we used a MLP with a single hidden layer with batch normalisation (Ioffe & Szegedy, 2015) and a single neuron as expert networks in the AME models. Each expert network received the gene expression at one gene locus as its input. For the FNN baseline, we chose the matching hyperparameters and architecture (100 neurons, 1 hidden layer). We optimised the neural networks with a learning rate of 0.0001, a batch size of 8 and an early stopping patience of 12 epochs. We trained each model on 35 random initialisations.

Pre- and Postprocessing. We standardised the input gene expression levels to have zero mean and unit variance. We applied the normalising transform (eq. 13) to DeepLIFT, LIME, SHAP and RF.

Metrics. We compared the error rates on the test set to assess predictive performance. In order to determine whether the associations identified by the various methods are consistent with those reported by domain experts, we counted the number of gene-cancer associations that were substantiated by literature evidence in the top 10 genes by average importance on the test set (Recall@10). We performed a literature search to determine which associations have previously been reported by domain experts. Appendix 4.7.H contains references and details of the literature search.

Results. We found that the AME(α=0.05) based its decisions primarily on a small number of highly predictive genes for the different types of cancer (Figure 4.4), and that literature evidence substantiated all of the top 10 links between respective cancer type and gene locus it reported (Table 4.3). AME(α=0) collapsed to assign an attention factor of 1 to one gene locus and 0 to all others for each cancer type - only reporting five non-zero importance scores. DeepLIFT, LIME and RF had difficulties discerning the important from the uninformative genes and assigned moderate levels of importance to many gene loci. DeepLIFT, LIME and SHAP were conflicted about which genes were relevant for which cancer with several of their top genes having high importance scores for multiple cancers. In contrast, AME(α=0.05) clearly distinguished both between cancers and important and uninformative genes. RF achieved a similar performance in terms of Recall@10 as AME(α=0.05). However, RFs can only produce an average set of importance scores for the whole training set. AMEs learn to accurately assign feature importance for individual samples, and can therefore explain every single prediction they make. On the test set, the mean$\pm$standard deviation of error rates across 35 runs of FNN, AME(α=0) and AME(α=0.05) were $1.10\pm0.005\%$, $4.86\pm0.093\%$, and $2.16\pm0.009\%$, respectively.



(a) AME(α=0.05)

(b) AME(α=0)

(c) DeepLIFT

(d) LIME(k=400)

(e) SHAP(k=100)

(f) RF

Figure 4.4: The importance of specific genes (coloured bars) for distinguishing between multiple cancer types as measured by average assignment of attention factors a_i. We report the average attention factors over All and over the per-cancer subsets. The grey bars spanning through the subsets highlight the 10 most discriminatory genes by average attention over All. We bolded the names of those genes whose associations are substantiated by literature evidence.

4.6 Conclusion

We presented a new approach to estimating feature importance that is based on the idea of distributing the feature groups of interest among experts in a mixture of experts model. The mixture of experts uses attentive gates to assign attention factors to individual experts. We introduced a secondary Granger-causal objective that defines feature importance as the marginal contribution towards prediction accuracy to ensure that the assigned attention factors correlate with the importance of the experts' input features. We showed that AMEs (i) compare favourably to several state-of-the-art methods in importance estimation accuracy, (ii) are significantly faster than existing methods, and (iii) discover associations that are consistent with those reported by domain experts. In addition, we found that there was a trade-off between predictive performance and accurate importance estimation when optimising jointly for both, that training with the Granger-causal objective was crucial to obtain accurate estimates, and that a lower Granger-causal error correlated with a better expected importance estimation accuracy. AMEs are a fast and accurate alternative that may be used when model-agnostic feature importance estimation methods are prohibitively expensive to compute. We believe AMEs could therefore be a first step towards translating the strong performance of neural networks in many domains into a deeper understanding of the underlying data.

4.7 Supplementary Material

4.7.A Source Code

The source code for this work is available at:
`https://github.com/d909b/ame`.

4.7.B Implementation Details

We used TensorFlow (Abadi et al., 2016) and Keras (Chollet et al., 2015) to implement the neural networks. We optimised all neural networks with the Adam (Kingma & Ba, 2015) optimiser. An important implementation detail is that one must prevent gradient flow from the Granger-causal objective towards the error term $\Delta\varepsilon_{X,i}$ to prevent the network from optimising the error to match the attributions rather than the other way around. We used open source implementations for LIME[2], SHAP[3] and DeepLIFT[4].

4.7.C Experiment 1: Architectures and Hyperparameters

The binary classifier used to evaluate the change in log odds was a feed-forward neural network (FNN) that consisted of 2 fully-connected hidden layers each with 128 neurons. Each hidden layer in the FNN was followed by a batch normalisation (BN) (Ioffe & Szegedy, 2015) layer, dropout (Srivastava et al., 2014) and a rectified linear unit (ReLU) (Nair & Hinton, 2010) activation. All feature importance estimation methods were evaluated against the same FNN. The attentive mixtures of experts (AMEs) used a shared convolutional network with a kernel size of 2 and stride of 2 - i.e. non-overlapping 2x2 blocks of pixels - whose output was split block by block into a total of $14 * 14 = 196$ independent experts. We used a single output neuron followed by a sigmoid activation as architecture for the auxiliary predictors $\mathrm{P}_{\mathrm{aux},i}$ and $\mathrm{P}_{\mathrm{aux},c}$. The output neuron indicated whether or not a given input sample was the target digit class. We did not use BN or dropout in the AMEs. We used three different AMEs with α set to 0, 0.1, 0.03, and 0.01. We optimised the neural networks' binary cross entropy with a learning rate of 0.001 and a batch size of 100 for the FNN and 500 for the AMEs. The FNN achieved an accuracy of 99% for distinguishing between the source and the target digit on the test set. We applied LIME and SHAP to the AME(α=0.1) with $k = 10\,000$ neighbourhood samples. To speed up the computation, we

[2] `https://github.com/marcotcr/lime`; accessed 20th January 2018
[3] `https://github.com/slundberg/shap`; accessed 20th January 2018
[4] `https://github.com/kundajelab/deeplift`; accessed 20th January 2018

used super-pixel segmentation with $14 * 14 = 196$ segments for LIME and SHAP.

We were unable to include DeepLIFT in this experiment, because there currently exists no DeepLIFT propagation rule for attentive gating networks that would have enabled us to apply DeepLIFT to find the most important pixels for the AME(α=0.1). However, (Lundberg & Lee, 2017) presented a comparison on this task that includes LIME, SHAP and DeepLIFT on a different model.

4.7.D Experiment 1: Samples of Masked Digits

We present samples of masked digits for AME(α=0), AME(α=0.01), AME(α=0.03), AME(α=0.1) and SHAP in Figures 4.5, 4.6, 4.7, 4.8 and 4.9, respectively.

4.7.E Experiment 2: Dataset

For the chosen time frame, we collected data from the following open data sources:

Demand History. As the primary data stream, we used the monthly data on all revenues generated by reimbursed prescription items in 9455 general practices (GPs) in England, UK as released by the British National Health Service (NHS)[5]. To ensure a full history of the past demand for predictions, we removed those practices from the dataset that remained closed at any point during the prediction time frame. To remove the noise introduced by different prescription formulations and packagings, we aggregated all prescription items into their prescription item class as defined by their pharmaceutical code in the British National Formulary (BNF) reported by the NHS.

[5]https://data.gov.uk/dataset/prescribing-by-gp-practice-presentation-level; accessed 20th January 2018

For context, we added the monthly total revenue, the region the current practice belongs to and the practice's distance to that region's centre as additional input features. We used the territory level 3 (TL3) regions as defined by (OECD, 2017) to geographically subdivide England into 145 regions.

Online Search Interest. We used Google Trends[6] to retrieve the relative monthly online search interest for our time frame and regional code (GB-EN). In total, we queried online search interest for 109 779 medical terms from the comprehensive medical subject headings (MeSH) ontology (Lipscomb, 2000). As many of the medical terms in the MeSH ontology were not common in vernacular, ultimately only 9 225 (8.40%) out of the 109 779 medical terms had a search activity that was significant enough to return a result on Google Trends. We applied a principal component analysis (PCA) transform that explained more than 98% of the variance in the search interest data and reduced the feature vector to 23 dimensions.

Regional Weather. Weather impacts consumption across a wide variety of industries (Lazo et al., 2011). We therefore added monthly weather data, including rainfall, days of air frost, hours of sunshine and average temperature data, from 23 UK Met Office[7] stations to our dataset.

Regional Demographic, Economic and Labor Data. Demographic, economic and labor data in proximity to a practice could affect future prescription demand, as groups with different population-level profiles potentially require different types of care in different amounts. To further analyse the predictive potential of population-level factors, we added 45 yearly indicators on the demographic profile, 27 yearly indicators on the economic profile and 15 yearly indicators of the labor profile of each TL3 region from (OECD, 2017) to our dataset. We represented each of the three population-level feature groups of indicators as a distinct expert in our AME model. Analogous

[6]https://trends.google.com/; accessed 20th January 2018

[7]http://www.metoffice.gov.uk/public/weather/climate-historic/; accessed 20th January 2018

to the online search interest data, we applied PCA transforms that explained over 99% of the variance in the three feature groups.

MeSH Term Selection. We extracted all MeSH terms from the categories 'Diseases' (C), 'Drugs and Chemicals' (D) and 'Mental Disorders' (F03).

Choice of Time Frame. We chose the evaluated time frame because overlapping and continuous historical data was available from all listed data sources.

Closed Practices. For our purposes, we defined closed practices as those practices that have had no revenue for more than 3 continuous months throughout the 24 month timeframe that we analysed.

Table 4.4 shows the full list of input features per feature group used in the experiment.

4.7.F Experiment 2: Architectures

We attempted to keep the architectures of the different models as similar as possible to ensure performance differences were not due to factors other than the design choices introduced by AMEs.

FNN Architecture. The FNNs were multi-layer perceptrons (MLPs) that had as input all the features of the six feature groups (Table 4.4) collapsed over time, i.e. flattened into a feature vector of a single dimension. The MLP consisted of L hidden layers with N fully-connected neurons each, where L and N were hyperparameters chosen at random across the 35 training runs. Each hidden layer was followed by a BN layer, dropout and a ReLU activation.

RNN Architecture. The recurrent neural networks (RNNs) were architecturally equivalent to the FNNs except that the inputs were not collapsed

over time and that the first fully-connected hidden layer was replaced with a bidirectional long short-term memory layer (Graves & Schmidhuber, 2005; Graves et al., 2013). In addition, all features that had no temporal dimension were replicated at each time step to match the temporal data. Each fully-connected hidden layer was followed by a BN layer, dropout and a ReLU activation.

AME Architecture. The AMEs consisted of $M = 6$ expert networks. The expert networks were themselves equivalent to the FNNs used in the same experiment, i.e. L hidden layers with N fully-connected neurons each, where L and N were hyperparameters chosen at random across the 35 training runs. The inputs to the experts in the AME were the six sets of inputs corresponding to the semantic feature groups (Table 4.4). Each hidden layer was followed by a BN layer, dropout and a ReLU activation. We used single fully-connected output neurons without an activation as auxiliary predictors $\mathrm{P}_{\mathrm{aux},i}$ and $\mathrm{P}_{\mathrm{aux},c}$.

All features that were on a nominal scale were converted to vector representations using neural embeddings (Mikolov et al., 2013; Goldberg & Levy, 2014) before being processed by the first hidden layer in the network. Figure 4.10 provides a visual representation of the employed AME architecture.

4.7.G Experiment 3: Architectures

FNN Architecture. The FNNs were MLPs that received as input all 100 gene expression levels. The MLP consisted of a single hidden layer with 100 fully-connected neurons. The hidden layer was followed by a BN layer and a ReLU activation. We used a five-way fully-connected softmax as output representing the five different types of cancer.

AME Architecture. The AME consisted of $M = 100$ expert networks. Each expert network used 1 hidden layers with a single neuron connected to the

gene expression level input at a distinct gene locus. The hidden layers were followed by a BN layer and a ReLU activation. We used five-way fully-connected softmax output neurons corresponding to the five different types of cancer as auxiliary predictors $P_{aux,i}$ and $P_{aux,c}$.

4.7.H Experiment 3: Literature Review Methodology

We searched Google Scholar[8] and the National Institutes of Health (NIH) gene database[9] for reported associations between the gene name and cancer type using the query "*gene_name* AND *cancer_type*". We took a conservative approach towards assessing whether gene-cancer associations were substantiated by biomedical literature in order to avoid biases stemming from assessing the strength of literature evidence. We considered any gene-cancer link as substantiated by literature if there existed any works at the time of our search that suggested even the slightest association between the two. When a gene locus had high importances for multiple cancer types, we screened for all of them as potential associations and considered the link to be substantiated if prior reports existed for any of them. If a gene locus had high importances for all but one cancer type, we also screened for an association with the left-out cancer type, in case the model used the transcription level at that gene locus as negative evidence. Table 4.6 contains all gene-cancer links for the associations reported by each method that we found to be substantiated by literature evidence and the corresponding references to literature.

[8] https://scholar.google.com/; accessed 1st May 2018
[9] https://www.ncbi.nlm.nih.gov/gene/; accessed 1st May 2018

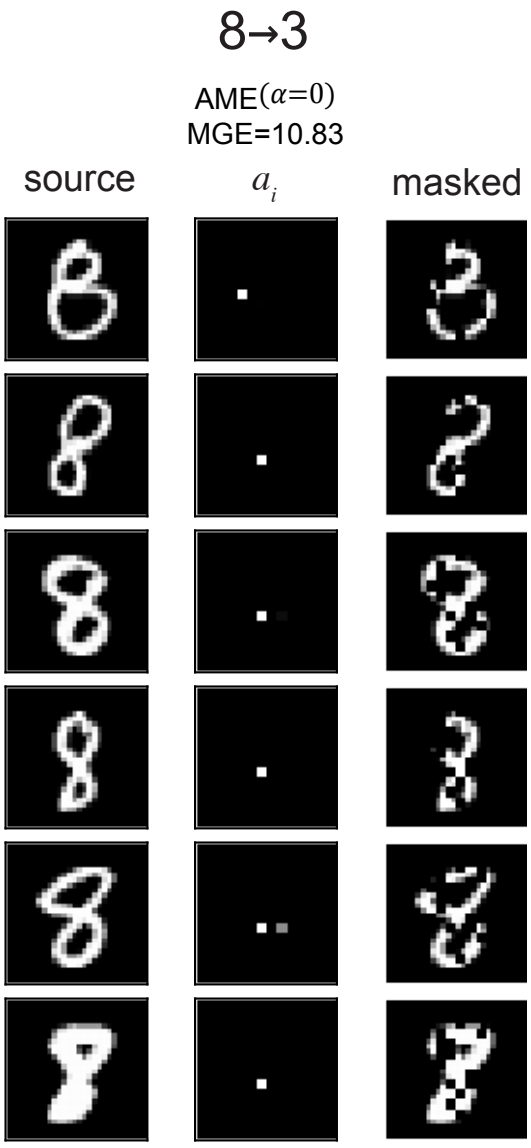


Figure 4.5: Samples of digits for the AME(α=0) from experiment 1.

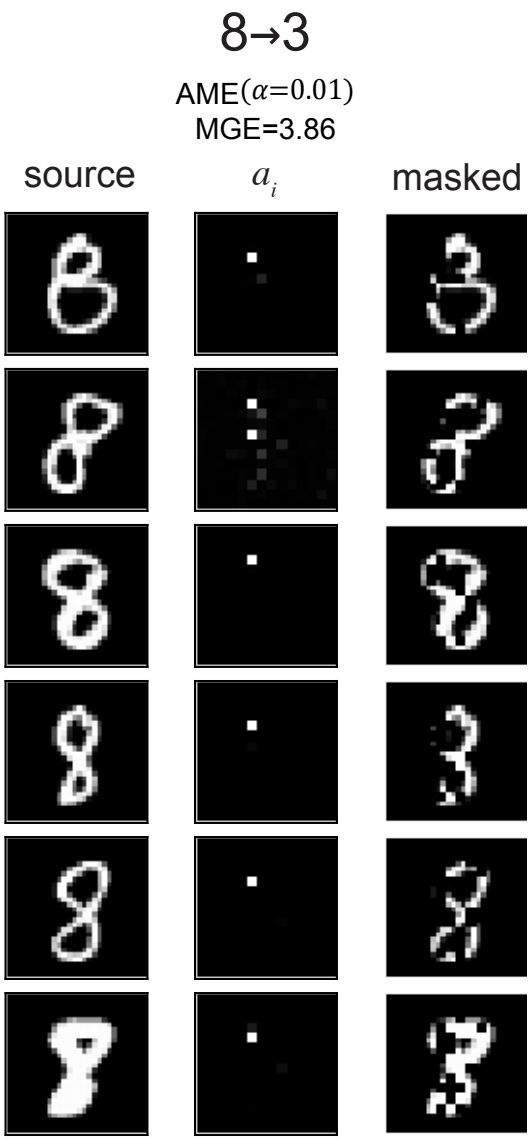


Figure 4.6: Samples of digits for the AME(α=0.01) from experiment 1.

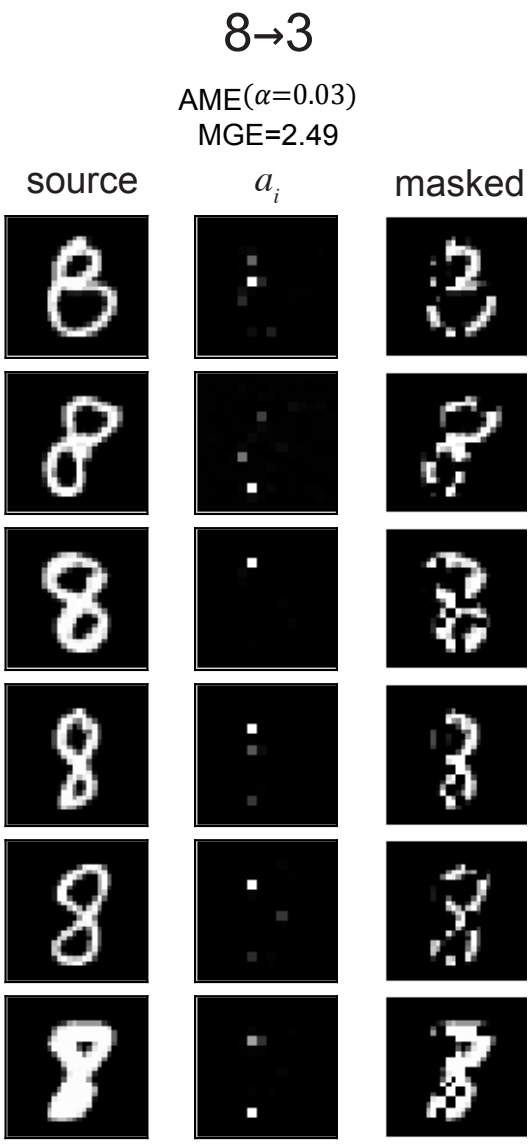


Figure 4.7: Samples of digits for the AME(α=0.03) from experiment 1.

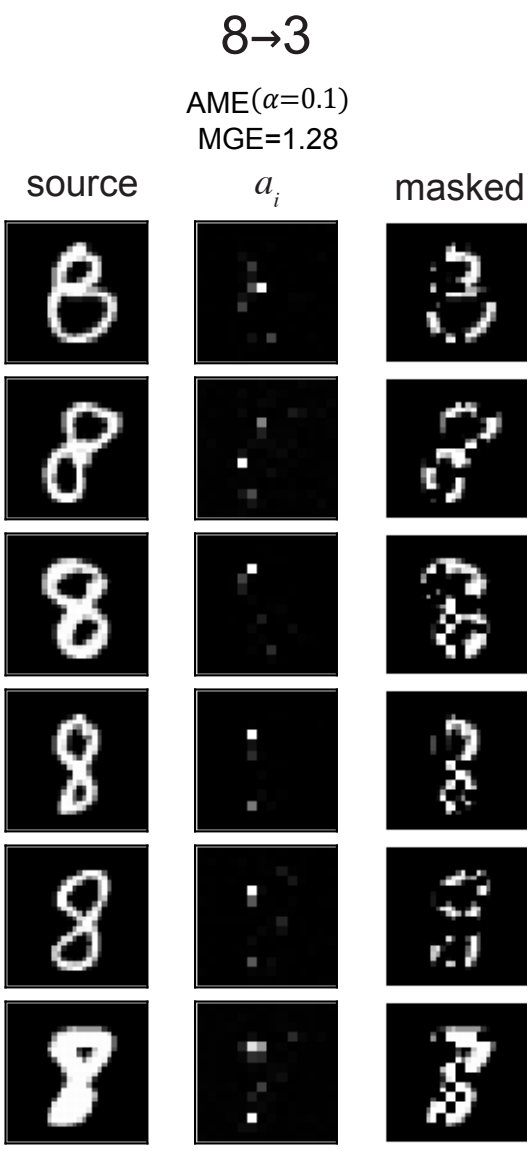


Figure 4.8: Samples of digits for the AME(α=0.1) from experiment 1.

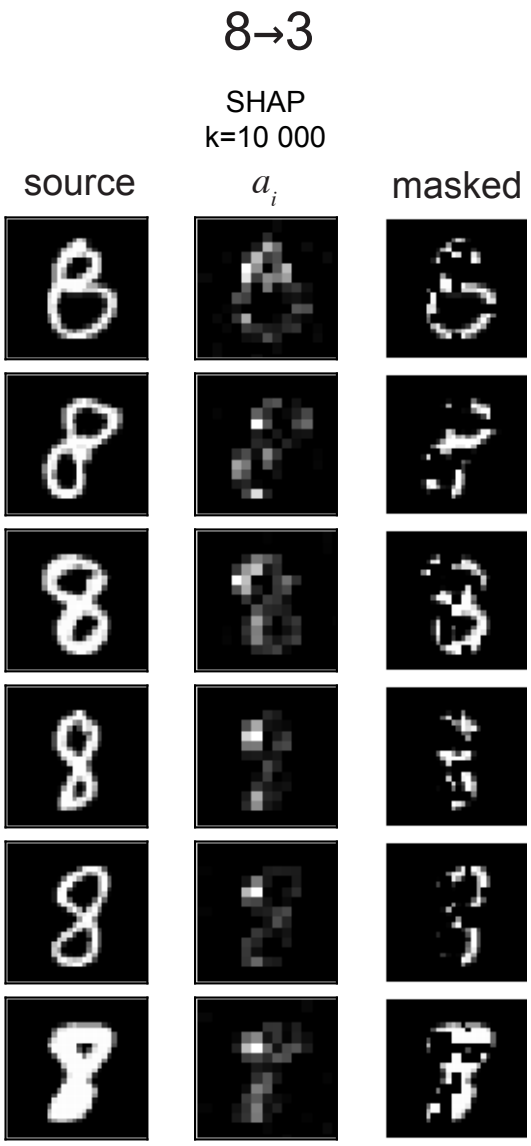


Figure 4.9: Samples of digits for SHAP(k=10 000) from experiment 1.

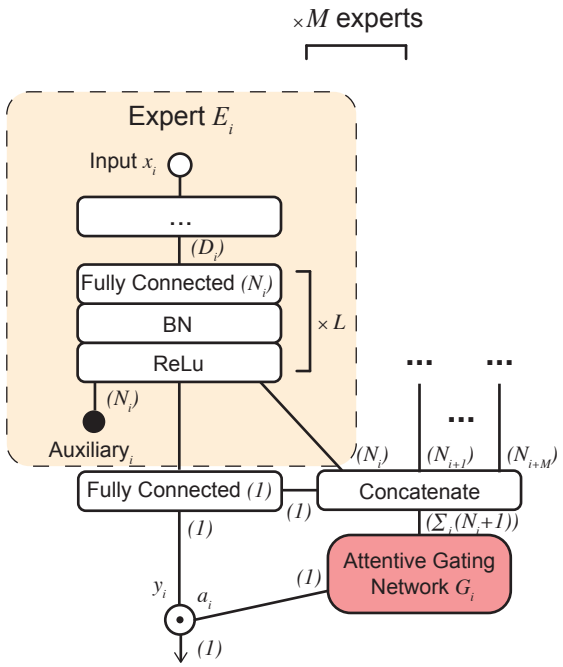


Figure 4.10: An illustration of the architecture of the AME presented in the medical prescription demand forecasting experiment. The AME consisted of $M = 6$ (architecturally equivalent) experts E_i that each operated on their own distinct input feature group x_i of dimensionality D_i after collapsing the temporal dimension (see Table 4.4 for the exhaustive list of inputs for each expert). The "..." layer indicates the input-dependant data transformation that collapsed the temporal dimension for those input features that had a temporal dimension. Each expert had its own respective number of units per hidden layer N_i with the hidden blocks repeated L times. Note that the evaluated feedforward neural networks (FNNs) used the same architecture as a single expert E_i of the AME that received all inputs x_i at once and did not have its output modulated by an attentive gating network G_i. The attentive gating networks consisted of a soft attention mechanism (Yang et al., 2016; Schwab et al., 2017) that attended to the combined hidden state h_c of the AME. For the evaluated recurrent neural networks (RNNs), we used the same architecture as for the FNNs but did not collapse the input features x in time and replaced the first fully connected hidden block with a bidirectional long short-term memory layer. The numbers in parentheses indicate the dimensionality of the connections between components of the AME.

Table 4.4: Full list of input features used in the multivariate models in the medical prescription demand forecasting experiment.

Input Feature	Dimensionality
(a) Demand History	
Monthly Prescription item revenue history	12×1
Min. and max. revenue	2
Monthly total revenues at practice	12×1
Month of forecast (one-hot)	12
TL3 Region (one-hot)	145
Prescription item type (embedding)	45
(b) Online Search Interest	
Monthly online search interest in 9 225 search terms (PCA)	12×23
Prescription item type (embedding)	45
(c) Regional Weather	
Monthly weather data	12×5
Distance to weather station	1
Prescription item type (embedding)	45
(d) Regional Demographic Data	
Distance to TL3 region	1
Prescription Item type (embedding)	45
Note: All of the following features were aggregated into a feature vector using PCA.	
Population, Number of people	1
Age-adjusted mortality rate, Deaths per 1 000 population	1
Crude youth death rate in age band 0 - 14 years, Deaths per 100 000 population	1
Crude death rate, Deaths per 100 000 population	1
Population in age band 0 - 4 years	1
Population in age band 5 - 9 years	1
Population in age band 10 - 14 years	1
Population in age band 15 - 19 years	1
Population in age band 20 - 24 years	1
Population in age band 25 - 29 years	1
Population in age band 30 - 34 years	1
Population in age band 35 - 39 years	1
Population in age band 40 - 44 years	1
Population in age band 45 - 49 years	1
Population in age band 50 - 54 years	1
Population in age band 55 - 59 years	1
Population in age band 60 - 64 years	1
Population in age band 65 - 69 years	1
Population in age band 70 - 74 years	1
Population in age band 75 - 79 years	1
Population in age band 80 - max. years	1
Population in age band 0 - 14 years	1
Population in age band 15 - 64 years	1
Population in age band 65 - max. years	1
Deaths, age 0 - 4 years	1
Deaths, age 5 - 9 years	1
Deaths, age 10 - 14 years	1
Deaths, age 15 - 19 years	1
Deaths, age 20 - 24 years	1
Deaths, age 25 - 29 years	1
Deaths, age 30 - 34 years	1
Deaths, age 35 - 39 years	1
Deaths, age 40 - 44 years	1
Deaths, age 45 - 49 years	1
Deaths, age 50 - 54 years	1
Deaths, age 55 - 59 years	1
Deaths, age 60 - 64 years	1
Deaths, age 65 - 69 years	1
Deaths, age 70 - 74 years	1
Deaths, age 75 - 79 years	1
Deaths, age 80 - max. years	1
Deaths, age 0 - 14 years	1
Deaths, age 15 - 64 years	1
Deaths, age 65 - max. years	1
Deaths, all ages	1

Table 4.5: Full list of input features used in the multivariate models in the medical prescription demand forecasting experiment. (cont.)

Input Feature	Dimensionality
(e) Regional Economic Data	
Distance to TL3 region	1
Prescription item type (embedding)	45
Note: All of the following features were aggregated into a feature vector using PCA.	
Regional Gross Domestic Product, Pound Sterling (GBP)	1
Disposable Household Income, GBP	1
Primary income of private Households, GBP	1
Regional employment, Number of people	1
Employment in information and communication, Number of people	1
Employment in financial and insurance, Number of people	1
Employment in real estate, Number of people	1
Employment in manufacturing, Number of people	1
Employment in construction, Number of people	1
Employment in agriculture, forestry and fishing, Number of people	1
Employment in trade, repairs, transport, accommodation, food service, Num. of people	1
Employment in scientific, technical, administrative service, Num. of people	1
Employment in administration, education, human health, Num. of people	1
Employment in industry, including energy, Number of people	1
Employment in other services, Number of people	1
Regional Gross Value Added (GVA), GBP	1
GVA in information and communication, GBP	1
GVA in financial and insurance, GBP	1
GVA in real estate, GBP	1
GVA in manufacturing, GBP	1
GVA in construction, GBP	1
GVA in agriculture, GBP	1
GVA in distributive trade, repairs, transport, accommodation, food service, GBP	1
GVA in professional, scientific, technical, administrative, support service, GBP	1
GVA in public administration, education, human health, GBP	1
GVA in industry, including energy, GBP	1
GVA in other services, GBP	1
(f) Regional Labor Data	
Distance to TL3 region	1
Prescription item type (embedding)	45
Note: All of the following features were aggregated into a feature vector using a PCA.	
Unemployed, Number of people	1
Unemployment rate, % unemployed over labour force age 15+	1
Unemployment rate, age 15+, Growth index	1
Unemployment rate, age 15+, Growth index (2007=100)	1
Working age population, age 15-64, Number of people	1
Working age population, age 15+, Number of people	1
Labor force, age 15+, Number of people	1
Participation rate, age 15+, % labour force age 15+ over age 15+	1
Participation rate, age 15+, % labour force age 15+ over age 15+, Growth (2001=100)	1
Participation rate, age 15+, % labour force age 15+ over age 15+, Growth (2007=100)	1
Employment in age band 15 - max. years, Number of people	1
Employment in age band 15 - 64 years, Number of people	1
Employment rate, % employment age 15 - 64 over working age 15 - 64	1
Employment rate, % employment age 15 - max. over working age 15 - max.	1
Employment rate growth, 15 years old and over, Index (2001=100)	1
Employment rate growth, 15 years old and over, Index (2007=100)	1

Table 4.6: The gene-cancer links for the associations reported by each method that we found to be substantiated by literature evidence and the corresponding references to literature. The top 10 gene-cancer links are in order (top to bottom) of average assigned importance over all cancer types. The cancer types for which the gene locus had the highest average importances are listed in parentheses next to the gene names. × indicates that no prior reports for this gene-cancer association were found.

Rank	AME (10/10)
1	100133144 (LUAD) (Van Ree et al., 2010)
2	729884 (KIRC) (Aburatani et al., 2015)
3	90288 (PRAD) (Fagerberg et al., 2014; Cao et al., 2013)
4	AADAT (LUAD) (Brenner et al., 2013)
5	AAGAB (BRCA) (Wang et al., 2017)
6	ABCA13 (BRCA) (Hlaváč et al., 2013; Krøigård et al., 2018)
7	ABCB6 (BRCA) (Park et al., 2006)
8	ABCB9 (BRCA) (Gong et al., 2016)
9	ABCC9 (COAD) (Jansová et al., 2006)
10	ABCD2 (COAD) (Hlavata et al., 2012)

Rank	RF (10/10)
1	553137 (KIRC) (Yeatman et al., 2015)
2	A1BG (KIRC/COAD/LUAD) (Liu et al., 2012; Zhang et al., 2015)
3	AADAT (BRCA/PRAD) (Brenner et al., 2013)
4	AAGAB (BRCA/KIRC) (Wang et al., 2017)
5	ABCA12 (BRCA) (Park et al., 2006; Hlaváč et al., 2013)
6	ABCA3 (COAD/LUAD) (Szakács et al., 2004; Overbeck et al., 2013)
7	ABCC11 (BRCA/LUAD) (Yamada et al., 2013)
8	ABCC3 (BRCA/LUAD) (Partanen et al., 2012)
9	ABCC4 (BRCA/PRAD) (Low et al., 2009)
10	ABCC6P1 (BRCA/KIRC) (Piehler et al., 2008; Kringen et al., 2012)

Rank	SHAP (8/10)
1	553137 (KIRC/COAD) (Yeatman et al., 2015)
2	729884 (KIRC) (Aburatani et al., 2015)
3	90288 (COAD/PRAD) ×
4	A1BG (BRCA) (Smith et al., 2008)
5	A1CF (BRCA/COAD) ×
6	AADAT (LUAD) (Brenner et al., 2013)
7	ABCA3 (LUAD) (Szakács et al., 2004; Overbeck et al., 2013)
8	ABCB1 (LUAD) (Han et al., 2007; Gervasini et al., 2006)
9	ABCC3 (BRCA/PRAD) (Partanen et al., 2012)
10	ABCC9 (COAD) (Jansová et al., 2006)

Rank	LIME (8/10)
1	553137 (KIRC/COAD) (Yeatman et al., 2015)
2	729884 (LUAD/KIRC) (Aburatani et al., 2015)
3	A1BG (BRCA) (Smith et al., 2008)
4	A1CF (BRCA/COAD) ×
5	AASS (LUAD/BRCA/COAD) ×
6	ABCB9 (BRCA) (Gong et al., 2016)
7	ABCC11 (BRCA/LUAD) (Yamada et al., 2013)
8	ABCC3 (BRCA/COAD/LUAD) (Partanen et al., 2012)
9	ABCC4 (BRCA) (Low et al., 2009)
10	ABCC6P1 (BRCA/PRAD) (Piehler et al., 2008; Kringen et al., 2012)

Rank	DeepLIFT (7/10)
1	553137 (KIRC) (Yeatman et al., 2015)
2	729884 (LUAD) (Aburatani et al., 2015)
3	A1BG (BRCA) (Smith et al., 2008)
4	A1CF (COAD) ×
5	AADAC (LUAD) ×
6	AASS (LUAD/BRCA/COAD) ×
7	ABCC11 (BRCA) (Yamada et al., 2013)
8	ABCC4 (BRCA) (Low et al., 2009)
9	ABCC6P1 (BRCA/PRAD) (Piehler et al., 2008; Kringen et al., 2012)
10	ABCC8 (BRCA) (Lehman et al., 2008; Soucek et al., 2015; Kim et al., 2017)

Rank	Attention (2/10)
1	100134869 (BRCA) (Van Ree et al., 2010)
2	26823 (KIRC) ×
3	AADACL4 (COAD) ×
4	AARS2 (LUAD) (Jiang et al., 2017)
5	AASDHPPT (PRAD) ×
6	×
7	×
8	×
9	×
10	×

This page was intentionally left blank.

Chapter 5

Learning to Estimate Individual Dose-Response

"Experience may teach us what is, but never that it cannot be otherwise." - Immanuel Kant

Estimating what would be an individual's potential response to varying levels of exposure to a treatment is of high practical relevance for several important fields, such as healthcare, economics and public policy. However, existing methods for learning to estimate counterfactual outcomes from observational data are either focused on estimating average dose-response curves, or limited to settings with only two treatments that do not have an associated dosage parameter. Here, we present a novel machine-learning approach towards learning counterfactual representations for estimating individual dose-response curves for any number of treatments with continuous dosage parameters with neural networks. Building on the established potential outcomes framework, we introduce performance metrics, model selection criteria, model architectures, and open benchmarks for estimating individual dose-response curves. Our experiments show that the methods developed in this work set a new state-of-the-art in estimating individual dose-response.

5.1 Introduction

Estimating dose-response curves from observational data is an important problem in many domains. In medicine, for example, we would be interested in using data of people that have been treated in the past to predict

which treatments and associated dosages would lead to better outcomes for new patients (Imbens, 2000). This question is, at its core, a counterfactual one, i.e. we are interested in predicting what *would have happened if* we were to give a patient a specific treatment at a specific dosage in a given situation.

Answering such counterfactual questions is a challenging task that requires either further assumptions about the underlying data-generating process or prospective interventional experiments, such as randomised controlled trials (RCTs) (Stone, 1993; Pearl, 2009; Peters et al., 2017). However, performing prospective experiments is expensive, time-consuming, and, in many cases, ethically not justifiable (Schafer, 1982). Two aspects make estimating counterfactual outcomes from observational data alone difficult (Yoon et al., 2018; Schwab et al., 2018b): Firstly, we only observe the factual outcome and never the counterfactual outcomes that would potentially have happened had we chosen a different treatment option. In medicine, for example, we only observe the outcome of giving a patient a specific treatment at a specific dosage, but we never observe what would have happened if the patient was instead given a potential alternative treatment or a different dosage of the same treatment. Secondly, treatments are typically not assigned at random in observational data. In the medical setting, physicians take a range of factors, such as the patient's expected response to the treatment, into account when choosing a treatment option. Due to this treatment assignment bias, the treated population may differ significantly from the general population. A supervised model naïvely trained to minimise the factual error would overfit to the properties of the treated group, and therefore not generalise to the entire population.

To address these problems, we introduce a novel methodology for training neural networks for counterfactual inference that extends to any number of treatments with continuous dosage parameters. In order to control for the biased assignment of treatments in observational data, we combine our method with a variety of regularisation schemes originally developed for the discrete treatment setting, such as distribution matching (Johansson et al., 2016; Shalit et al., 2017), propensity dropout (PD) (Alaa et al., 2017), and

matching on balancing scores (Rosenbaum & Rubin, 1983; Ho et al., 2007; Schwab et al., 2018b). In addition, we devise performance metrics, model selection criteria and open benchmarks for estimating individual dose-response curves. Our experiments demonstrate that the methods developed in this work set a new state-of-the-art in inferring individual dose-response curves.

Contributions. This chapter contains the following contributions:

- We introduce a novel methodology for training neural networks for counterfactual inference that, in contrast to existing methods, is suitable for estimating counterfactual outcomes for any number of treatment options with associated exposure parameters.
- We develop performance metrics, model selection criteria, model architectures, and open benchmarks for estimating individual dose-response curves.
- We extend state-of-the-art methods for counterfactual inference for two non-parametric treatment options to the multiple parametric treatment options setting.
- We perform extensive experiments that show that our method sets a new state-of-the-art in inferring individual dose-response curves from observational data across several challenging datasets.

5.2 Related Work

Background. Causal analysis of treatment effects with rigorous experiments is, in many domains, an essential tool for validating interventions. In medicine, prospective experiments, such as RCTs, are the de facto gold standard to evaluate whether a given treatment is efficacious in treating a specific indication across a population (Carpenter, 2014; Bothwell et al., 2016). However, performing prospective experiments is expensive, time-consuming, and often not possible for ethical reasons (Schafer, 1982; Schwab

et al., 2018b). Historically, there has therefore been considerable interest in developing methodologies for performing causal inference using readily available observational data (Granger, 1969; Angrist et al., 1996; Rosenbaum & Rubin, 1983; Robins et al., 2000; Pearl, 2009; Hernán & Robins, 2016; Lake et al., 2017). The naïve approach of training supervised models to minimise the observed factual error is in general not a suitable choice for counterfactual inference tasks due to treatment assignment bias and the inability to observe counterfactual outcomes. To address the shortcomings of unsupervised and supervised learning in this setting, several adaptations to established machine-learning methods that aim to enable the estimation of counterfactual outcomes from observational data have recently been proposed (Johansson et al., 2016; Shalit et al., 2017; Wager & Athey, 2017; Alaa & van der Schaar, 2017; Alaa et al., 2017; Louizos et al., 2017; Yoon et al., 2018; Schwab et al., 2018b). In this work, we build on several of these advances to develop a machine-learning approach for estimating individual dose-response with neural networks.

Estimating Individual Treatment Effects (ITE). [1] Matching methods (Ho et al., 2007) are among the most widely used approaches to causal inference from observational data. Matching methods estimate the counterfactual outcome of a sample X to a treatment t using the observed factual outcome of its nearest neighbours that have received t (Schwab et al., 2018b). Propensity score matching (PSM) (Rosenbaum & Rubin, 1983) combats the curse of dimensionality of matching directly on the covariates X by instead matching on the scalar probability $p(t|X)$ of receiving a treatment t given the covariates X (Schwab et al., 2018b). Another category of approaches uses adjusted regression models that receive both the covariates X and the treatment t as inputs (Schwab et al., 2018b). The simplest such model is Ordinary Least Squares (OLS), which may use either one model for all treatments, or a separate model for each treatment (Kallus, 2017; Schwab et al., 2018b). More complex models based on neural networks, like Treatment Agnostic Representation Networks (TARNETs), may be used to

[1] The ITE is sometimes also referred to as the conditional average treatment effect (CATE).

build non-linear regression models (Shalit et al., 2017; Schwab et al., 2018b). Estimators that combine a form of adjusted regression with a model for the exposure in a manner that makes them robust to misspecification of either are referred to as doubly robust (Funk et al., 2011; Schwab et al., 2018b). In addition to OLS and neural networks, tree-based estimators, such as Bayesian Additive Regression Trees (BART) (Chipman et al., 2010; Chipman & McCulloch, 2016) and Causal Forests (CF) (Wager & Athey, 2017), and distribution modelling methods, such as Causal Multi-task Gaussian Processes (CMGP) (Alaa & van der Schaar, 2017), Causal Effect Variational Autoencoders (CEVAEs) (Louizos et al., 2017), and Generative Adversarial Nets for inference of Individualised Treatment Effects (GANITE) (Yoon et al., 2018), have also been proposed for ITE estimation[2] (Schwab et al., 2018b). Other approaches, such as balancing neural networks (BNNs) (Johansson et al., 2016) and counterfactual regression networks (CFRNET) (Shalit et al., 2017), attempt to achieve balanced covariate distributions across treatment groups by explicitly minimising the empirical discrepancy distance between treatment groups using metrics such as the Wasserstein distance (Cuturi, 2013; Schwab et al., 2018b). Most of the works mentioned above focus on the simplest setting with two available treatment options without associated dosage parameters. A notable exception is the generalised propensity score (GPS) (Imbens, 2000) that extends the propensity score to treatments with continuous dosages.

In contrast to existing methods, we present the first machine-learning approach to learn to estimate individual dose-response curves for multiple available treatments with a continuous dosage parameter from observational data with neural networks. We additionally extend several known regularisation schemes for counterfactual inference to address the treatment assignment bias in observational data. To facilitate future research in this important area, we introduce performance metrics, model selection criteria, and open benchmarks. We believe this work could be particularly important for applications in precision medicine, where the current state-of-the-art of

[2]See (Knaus et al., 2018) and (Schwab et al., 2018b) for empirical comparisons of large-numbers of machine-learning methods for ITE estimation for two and more available treatment options.

estimating the average dose response across the entire population does not take into account individual differences, even though large differences in dose-response between individuals are well-documented for many diseases (Zaske et al., 1982; Oldenhof et al., 1988; Campbell et al., 2007).

5.3 Methodology

Problem Statement. We consider a setting in which we are given N observed samples X with p pre-treatment covariates x_i and $i \in [0 \,..\, p-1]$. For each sample, the potential outcomes $y_{n,t}(s_t)$ are the response of the nth sample to a treatment t out of the set of k available treatment options $T = \{0, ..., k-1\}$ applied at a dosage $s_t \in \{s_t \in \mathbb{R}, a_t > 0 \mid a_t \leq s \leq b_t\}$, where a_t and b_t are the minimum and maximum dosage for treatment t, respectively. The set of treatments T can have two or more available treatment options. As training data, we receive factual samples X and their observed outcomes $y_{n,f}(s_f)$ after applying a specific observed treatment f at dosage s_f. Using the training data with factual outcomes, we wish to train a predictive model to produce accurate estimates $\hat{y}_t(n, s)$ of the potential outcomes across the entire range of s for all available treatment options t. We refer to the range of potential outcomes $y_{n,t}(s)$ across s as the *individual dose-response curve* of the nth sample. This setting is a direct extension of the Rubin-Neyman potential outcomes framework (Rubin, 2005).

Assumptions. Following (Imbens, 2000; Lechner, 2001), we assume unconfoundedness, which consists of three key parts: (1) Conditional Independence Assumption: The assignment to treatment t is independent of the outcome y_t given the pre-treatment covariates X, (2) Common Support Assumption: For all values of X, it must be possible to observe all treatment options with a probability greater than 0, and (3) Stable Unit Treatment Value Assumption: The observed outcome of any one unit must be unaffected by the assignments of treatments to other units. In addition, we assume smoothness, i.e. that units with similar covariates x_i have similar outcomes y, both for model training and selection.

Metrics. To enable a meaningful comparison of models in the presented setting, we use metrics that cover several desirable aspects of models trained for estimating individual dose-response curves. Our proposed metrics respectively aim to measure a predictive model's ability (1) to recover the dose-response curve across the entire range of dosage values, (2) to determine the optimal dosage point for each treatment, and (3) to deduce the optimal treatment policy overall, including selection of the right treatment and dosage point, for each individual case. To measure to what degree a model covers the entire range of individual dose-response curves, we use the mean integrated square error[3] (MISE) between the true dose-response y and the predicted dose-response $\hat{y}$ as estimated by the model over N samples, all treatments T, and the entire range $[a_t, b_t]$ of dosages s.

$$\text{MISE} = \frac{1}{N}\frac{1}{|T|}\sum_{t\in T}\sum_{n=1}^{N}\int_{s=a_t}^{b_t}\Big(y_{n,t}(s) - \hat{y}_{n,t}(s)\Big)^2 ds \tag{5.1}$$

To further measure a model's ability to determine the optimal dosage point for each individual case, we calculate the mean dosage policy error (DPE). The mean dosage policy error is the mean squared error in outcome y associated with using the estimated optimal dosage point $\hat{s}_t^*$ according to the predictive model to determine the *true* optimal dosage point s_t^* over N samples and all treatments T.

$$\text{DPE} = \frac{1}{N}\frac{1}{|T|}\sum_{t\in T}\sum_{n=1}^{N}\Big(y_{n,t}(s_t^*) - y_{n,t}(\hat{s}_t^*)\Big)^2 \tag{5.2}$$

where s_t^* and $\hat{s}_t^*$ are the optimal dosage point according to the true dose-response curve and the estimated dose-response curve, respectively.

$$s_t^* = \underset{s\in[a_t,b_t]}{\arg\max}\, y_{n,t}(s) \tag{5.3}$$

$$\hat{s}_t^* = \underset{s\in[a_t,b_t]}{\arg\max}\, \hat{y}_{n,t}(s) \tag{5.4}$$

Finally, the policy error (PE) measures a model's ability to determine the optimal treatment policy for individual cases, i.e. how much worse the outcome would be when using the estimated best optimal treatment option

[3] A normalised version of this metric has been used in Silva (2016).

as opposed to the *true* optimal treatment option and dosage.

$$\text{PE} = \frac{1}{N} \sum_{n=1}^{N} \left(y_{n,t^*}(s^*_{t^*}) - y_{n,\hat{t}^*}(\hat{s}^*_{\hat{t}^*}) \right)^2 \tag{5.5}$$

where

$$t^* = \arg\max_{t \in T} y_{n,t}(s^*_t) \tag{5.6}$$

$$\hat{t}^* = \arg\max_{t \in T} \hat{y}_{n,t}(\hat{s}^*_t) \tag{5.7}$$

are the optimal treatment option according to the ground truth y and the predictive model, respectively. Considering the DPE and PE alongside the MISE is important to comprehensively evaluate models for counterfactual inference. For example, a model that accurately recovers dose response curves outside the regions containing the optimal response would achieve a respectable MISE but would not be a good model for determining the treatment and dosage choices that lead to the best outcome for a given unit. By considering multiple metrics, we can ensure that predictive models are capable both in recovering the entire dose-response as well as in selecting the best treatment and dosage choices. We note that, in general, we can not calculate the MISE, DPE or PE without knowledge of the outcome-generating process, since the true dose-response function $y_{n,t}(s)$ is unknown.

Model Architecture. Model structure plays an important role in learning representations for counterfactual inference with neural networks (Shalit et al., 2017; Schwab et al., 2018b; Alaa & Schaar, 2018). A particularly challenging aspect of training neural networks for counterfactual inference is that the influence of the treatment indicator variable t may be lost in high-dimensional hidden representations (Shalit et al., 2017). To address this problem for the setting of two available treatments without dosage parameters, Shalit et al. (2017) proposed the TARNET architecture that uses a shared base network and separate head networks for both treatment options. In TARNETs, the head networks are only trained on samples that received the respective treatment. Schwab et al. (2018b) extended the TARNET architecture to the multiple treatment setting by using k

separate head networks, one for each treatment option. In the setting with multiple treatment options with associated dosage parameters, this problem is further compounded because we must maintain not only the influence of t on the hidden representations throughout the network, but also the influence of the continuous dosage parameter s. To ensure the influence of both t and s on hidden representations, we propose a hierarchical architecture for multiple treatments called dose response network (DRNet, Figure 5.1). DRNets ensure that the dosage parameter s maintains its influence by assigning a head to each of $E \in \mathbb{N}$ equally-sized dosage strata that subdivide the range of potential dosage parameters $[a_t, b_t]$. The hyperparameter E defines the trade-off between computational performance and the resolution $\frac{(b-a)}{E}$ at which the range of dosage values is partitioned. To further attenuate the influence of the dosage parameter s within the head layers, we additionally repeatedly append s to each hidden layer in the head layers. We motivate the proposed hierarchical structure with the effectiveness of the regress and compare approach to counterfactual inference (Kallus, 2017), where one builds a separate estimator for each available treatment option. Separate models for each treatment option suffer from data-sparsity, since only units that received each respective treatment can be used to train a per-treatment model and there may not be a large number of samples available for each treatment. DRNets alleviate the issue of data-sparsity by enabling information to be shared both across the entire range of dosages through the treatment layers and across treatments through the base layers.

Model Selection. Given multiple models, it is not trivial to decide which model would perform better at counterfactual tasks, since we in general do not have access to the true dose-response to calculate error metrics like the ones given above. We therefore use a nearest neighbour approximation of the MISE to perform model selection using held-out factual data that has not been used for training. We calculate the nearest neighbour approximation

NN-MISE of the MISE using:

$$\text{NN-MISE} = \frac{1}{N}\frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\int_{s=a_t}^{b_t}\Big(y_{\text{NN}(n),t}(s) - \hat{y}_{n,t}(s)\Big)^2 ds \tag{5.8}$$

where we substitute the true dose-response $y_{n,t}$ of the nth sample with the outcome $y_{\text{NN}(n),t}$ of an observed factual nearest neighbour of the nth sample at a dosage point s from the training set. Using the nearest neighbour approximation of the MISE, we are able to perform model selection without access to the true counterfactual outcomes y. Among others, nearest neighbour methods have also been proposed for model selection in the setting with two available treatments without dosages (Schuler et al., 2018).

Regularisation Schemes. DRNets can be combined with regularisation schemes developed to further address treatment assignment bias. To determine the utility of various regularisation schemes, we evaluated DRNets using distribution matching (Shalit et al., 2017), propensity dropout (Alaa et al., 2017), matching on the entire dataset (Ho et al., 2007), and on the batch level (Schwab et al., 2018b). We naïvely extended these regularisation schemes since neither of these methods were originally developed for the dose-response setting (Appendix 5.7.B).

5.4 Experiments

Our experiments aimed to answer the following questions:

(1) How does the performance of our proposed approach compare to state-of-the-art methods for estimating individual dose-response?

(2) How do varying choices of E influence counterfactual inference performance?

(3) How does increasing treatment assignment bias affect the performance of dose-response estimators?

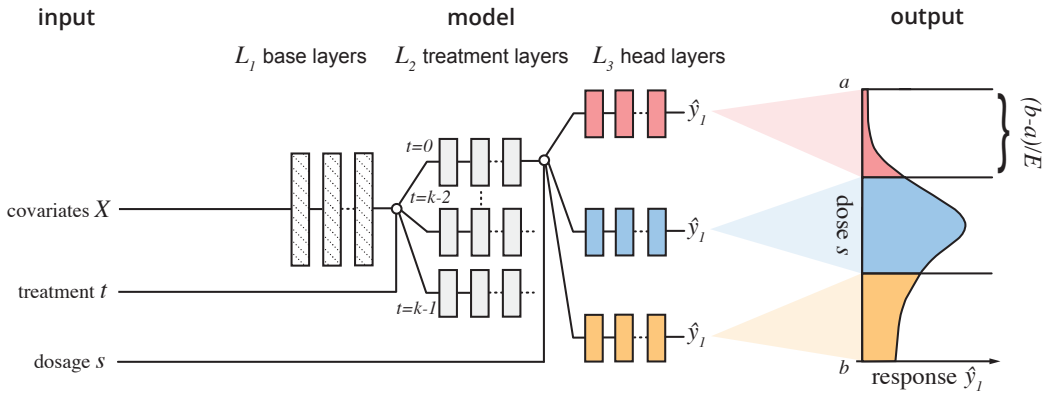


Figure 5.1: The dose response network (DRNet) architecture with shared base layers, k intermediary treatment layers, and $k*E$ heads for the multiple treatment setting with an associated dosage parameter s. The shared base layers are trained on all samples, and the treatment layers are only trained on samples from their respective treatment category. Each treatment layer is further subdivided into E head layers (only one set of $E = 3$ head layers for treatment $t = 0$ is shown above). Each head layer is assigned a dosage stratum that subdivides the range of potential dosages $[a_t, b_t]$ into E partitions of equal width $(b - a)/E$. The head layers each predict outcomes $\hat{y}_t(s)$ for a range of values of the dosage parameter s, and are only trained on samples that fall within their respective dosage stratum. The hierarchical structure of DRNets enables them to share common hidden representations across all samples (base layers), treatment options (treatment layers), and dosage strata (head layers) while maintaining the influence of both t and s on the hidden layers.

Table 5.1: Comparison of the benchmark datasets used in our experiments. We evaluate on three semi-synthetic datasets with varying numbers of treatments and samples.

Dataset	# Samples	# Features	# Treatments
News	5 000	2 870	2/4/8/16
MVICU	8 040	49	3
TCGA	9 659	20 531	3

Datasets. Using real-world data, we performed experiments on three semi-synthetic datasets with two and more treatment options to gain a better understanding of the empirical properties of our proposed approach. To cover a broad range of settings, we chose datasets with different outcome and treatment assignment functions, and varying numbers of samples, features and treatments (Table 5.1). All three datasets were randomly split into training (63%), validation (27%) and test sets (10%).

News. The News benchmark consisted of 5 000 randomly sampled news articles from the New York Times corpus[4] and was originally introduced as a benchmark for counterfactual inference in the setting with two treatment options - corresponding to different viewing devices - without an associated dosage parameter (Johansson et al., 2016). We extended the original dataset specification (Johansson et al., 2016; Schwab et al., 2018b) to enable the simulation of any number of treatments with associated dosage parameters. The samples X were news articles that consist of word counts $x_i \in \mathbb{N}$, outcomes $y_{s,t} \in \mathbb{R}$ that represent the reader's opinion of the news item, and a normalised dosage parameter $s_t \in (0, 1]$ that represents the viewer's reading time. There was a variable number of available treatment options t that corresponded to various devices that could be used to view the News items, e.g. smartphone, tablet, desktop, television or others (Johansson et al., 2016; Schwab et al., 2018b). We trained a topic model on the entire New York Times corpus to model that consumers prefer to read certain media items on specific viewing devices (Schwab et al., 2018b). We defined

[4] `https://archive.ics.uci.edu/ml/datasets/bag+of+words`; accessed 1st Feb 2019

$z(X)$ as the topic distribution of news item X, and randomly picked k topic space centroids z_t and $2k$ topic space centroids $z_{s_t,i}$ with $i \in 0, 1$ as prototypical news items (Schwab et al., 2018b). We assigned a random Gaussian outcome distribution with mean $\mu \sim \mathcal{N}(0.45, 0.15)$ and standard deviation $\sigma \sim \mathcal{N}(0.1, 0.05)$ to each centroid (Schwab et al., 2018b). For each sample, we drew ideal potential outcomes from that Gaussian outcome distribution $\tilde{y}_t \sim \mathcal{N}(\mu_t, \sigma_t) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 0.15)$ (Schwab et al., 2018b). The dose response $\tilde{y}_s$ was drawn from a distance-weighted mixture of two Gaussians $\tilde{y}_s \sim d_0\mathcal{N}(\mu_{s_t,0}, \sigma_{s_t,0}) + d_1\mathcal{N}(\mu_{s_t,1}, \sigma_{s_t,1})$ using topic space distances $d = \text{softmax}(\text{D}(z(X), z_{s_t,i}))$ and the Euclidean distance as distance metric D. We assigned the observed treatment t using $t|x \sim \text{Bern}(\text{softmax}(\kappa\tilde{y}_t\tilde{y}_s))$ with a treatment assignment bias coefficient κ and an exponentially distributed observed dosage s_t using $s_t \sim \text{Exp}(\beta)$ with $\beta = 0.25$. The true potential outcomes $y_{s,t} = C\tilde{y}_t\tilde{y}_s$ were the product of $\tilde{y}_t$ and $\tilde{y}_s$ scaled by a coefficient $C = 50$. We used four different variants of this dataset with $k = 2$, 4, 8, and 16 viewing devices, and $\kappa = 10$, 10, 10, and 7, respectively (Schwab et al., 2018b). Higher values of κ indicate a higher expected treatment assignment bias depending on $\tilde{y}_t\tilde{y}_s$, with $\kappa = 0$ indicating no assignment bias (Schwab et al., 2018b). A version of the News benchmark for the setting without associated dosage parameters appeared in Schwab et al. (2018b).

Mechanical Ventilation in the Intensive Care Unit (MVICU). The MVICU benchmark models patients' responses to different configurations of mechanical ventilation in the intensive care unit. The data was sourced from the publicly available MIMIC III database (Saeed et al., 2011). The samples X consisted of the last observed measurements x_i of various biosignals, including respiratory, cardiac and ventilation signals. The outcomes were arterial blood gas readings of the ratio of arterial oxygen partial pressure to fractional inspired oxygen PaO_2/FiO_2 which, at values lower than 300, are used as one of the clinical criteria for the diagnosis Acute Respiratory Distress Syndrome (ARDS) (Ferguson et al., 2012). We modelled a mechanical ventilator with $k = 3$ adjustable treatment parameters: (1) the fraction of inspired oxygen, (2) the positive end-expiratory pressure in

the lungs, and (3) tidal volume. To model the outcomes, we use the same procedure as for the News benchmark with a Gaussian outcome function and a mixture of Gaussian dose-response function, with the exception that we did not make use of topic models and instead performed the similarity comparisons D in covariate space. We used a treatment assignment bias $\kappa = 10$ and a scaling coefficient $C = 150$. Treatment dosages were drawn according to $s_t \sim \mathcal{N}(\mu_{\text{dose},t}, 0.1)$, where the distribution means were defined as $\mu_{\text{dose}} = (0.6, 0.65, 0.4)$ for each treatment. A preliminary version of the MVICU benchmark appeared in Linhardt (2018).

The Cancer Genomic Atlas (TCGA). The TCGA project collected gene expression data from various types of cancers in 9 659 individuals (Weinstein et al., 2013; Schwab et al., 2018b). There were $k = 3$ available clinical treatment options: (1) medication, (2) chemotherapy, and (3) surgery. We used a synthetic outcome function that simulated the risk of cancer recurrence after receiving either of the treatment options based on the real-world gene expression data (Schwab et al., 2018b). We standardised the gene expression data using the mean and standard deviations of gene expression at each gene locus for normal tissue in the training set (Schwab et al., 2018b). To model the outcomes, we followed the same approach as in the MVICU benchmark with similarity comparisons done in covariate space using the cosine similarity as distance metric D, and parameterised with $\kappa = 10$ and $C = 50$. Treatment dosages in the TCGA benchmark were drawn according to $s_t \sim \mathcal{N}(0.65, 0.1)$. A version of the TCGA benchmark for the setting without dosage parameters appeared in Schwab et al. (2018b).

5.4.1 Experimental Setup

Models. We evaluated DRNet, ablations, baselines, and all relevant state-of-the-art methods: k-nearest neighbours (kNN) (Ho et al., 2007), BART (Chipman et al., 2010; Chipman & McCulloch, 2016), CF (Wager & Athey, 2017), GANITE (Yoon et al., 2018), TARNET (Shalit et al., 2017), and GPS (Imbens, 2000) using the "causaldrf" package (Galagate, 2016).

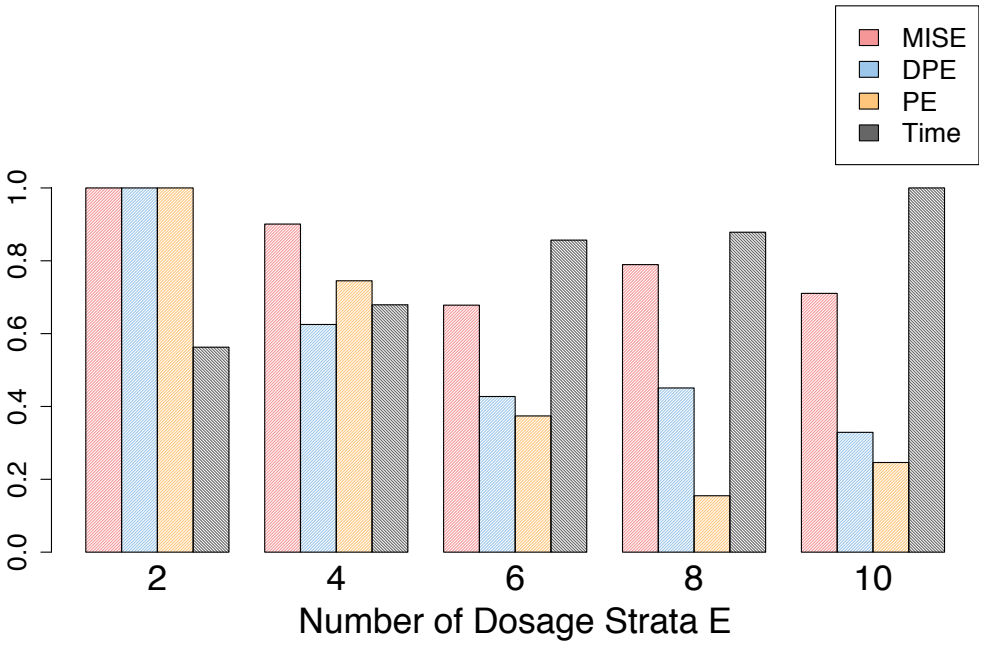


Figure 5.2: Analysis of the effect of choosing varying numbers of dosage strata E (x-axis) on MISE (red), DPE (blue), PE (orange) and Time needed for training and evaluation (black) as calculated on the MVICU benchmark. Metrics were normalised to the range $[0, 1]$. All other hyperparameters besides E were held equal.

We evaluated which regularisation strategy for learning counterfactual representations is most effective by training DRNets using a Wasserstein regulariser between treatment group distributions (+ Wasserstein) (Shalit et al., 2017), PD (+ PD) (Alaa et al., 2017), batch matching (+ PM) (Schwab et al., 2018b), and matching the entire training set as a preprocessing step (Ho et al., 2011) using the PM algorithm (+ PSM_{PM}) (Schwab et al., 2018b). To determine whether the DRNet architecture is more effective than its alternatives at learning representations for counterfactual inference in the presented setting, we also evaluated (1) a multi-layer perceptron (MLP) that received the treatment index t and dosage s as additional inputs, and (2) a TARNET for multiple treatments that received the dosage s as an extra input (TARNET) (Johansson et al., 2016; Schwab et al., 2018b) with all other hyperparameters beside the architecture held equal. As a final ablation of DRNet, we tested whether appending the dosage parameter s to each hidden layer in the head networks is effective by also training DRNets that only receive the dosage parameter once in the first hidden layer of the head network (- Repeat). We naïvely extended CF, GANITE and BART by adding the dosage as an additional input covariate, because they were not designed for treatments with dosages.

Hyperparameters. To ensure a fair comparison of the tested models, we took a systematic approach to hyperparameter search. Each model was given exactly the same number of hyperparameter optimisation runs with hyperparameters chosen at random from predefined hyperparameter ranges (Appendix 5.7.C). We chose a fixed hyperparameter optimisation budget to ensure that all methods received the same degree of hyperparameter optimisation. We used 5 hyperparameter optimisation runs for each model on TCGA and 10 on all other benchmarks. Furthermore, we used the same random seed for each model, i.e. all models were evaluated on exactly the same sets of hyperparameter configurations. After computing the hyperparameter runs, we chose the best model based on the validation set NN-MISE. This setup ensures that each model received the same degree of hyperparameter optimisation. For all DRNets and ablations, we used $E = 5$ dosage strata with the exception of those presented in Figure 5.2.

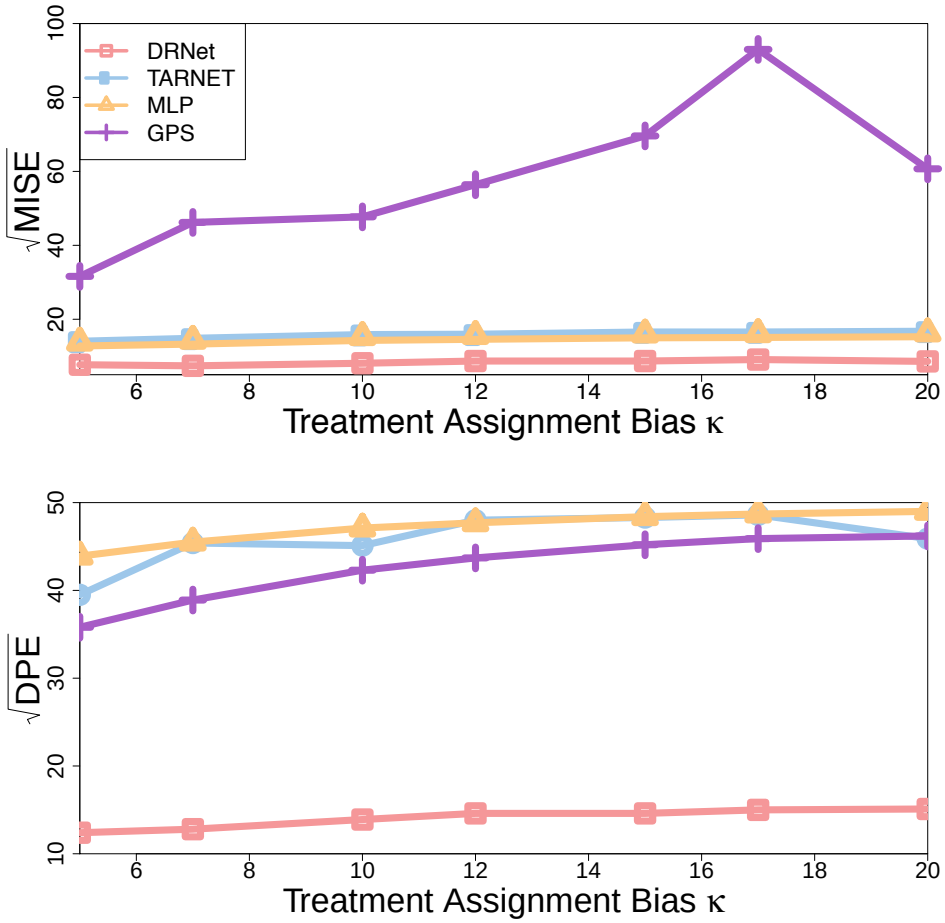


Figure 5.3: Comparison of DRNet (red), TARNET (blue), MLP (yellow) and GPS (purple) in terms of their $\sqrt{\text{MISE}}$ (top) and $\sqrt{\text{DPE}}$ (bottom) for varying levels of treatment assignment bias κ (x-axis) on News-2. DRNet performs better than other methods across the entire evaluated range of treatment assignment bias values, and is more robust to increasing levels of κ. Note that differences in the relative rankings of methods are possible across the metrics because a method could be, in relative terms, good at recovering the optimal dosage, even if it is not able to accurately recover the entire dose-response curve. This would be the case, for example, if the model is more accurate around values that are closer to the optimal dosage choice. Such a setting is common, since more training data is usually available for values closer to the optimal dosage because treatment assignments in observational data are often biased towards better outcomes.

Metrics. For each dataset and model, we calculated the $\sqrt{\text{MISE}}$, $\sqrt{\text{DPE}}$, and $\sqrt{\text{PE}}$. We used Romberg integration with 64 equally spaced samples from $y_{n,t}$ and $\hat{y}_{n,t}$ to compute the inner integral over the range of dosage parameters necessary for the MISE metric. To compute the optimal dosage points and treatment options in the DPE and PE, we used Sequential Least Squares Programming (SLSQP) to determine the respective maxima of $y_{n,t}(s)$ and $\hat{y}_{n,t}(s)$ numerically.

5.5 Results and Discussion

Counterfactual Inference. In order to evaluate the relative performances of the various methods across a wide range of settings, we compared the MISE of the listed models for counterfactual inference on the News-2/4/8/16, MVICU and TCGA benchmarks (Table 5.2; DPE and PE in Tables 5.3 and 5.4). Across the benchmarks, we found that DRNets outperformed all existing state-of-the-art methods in terms of MISE. We also found that DRNets that used additional regularisation strategies outperformed vanilla DRNets on News-2, News-4, News-8 and News-16. However, on MVICU and TCGA, DRNets that used additional regularisation performed similarly as standard DRNets. Where regularisation was effective, Wasserstein regularisation between treatment groups (+ Wasserstein) and batch matching (+ PM) were generally slightly more effective than PSM_{PM} and PD. In addition, not repeating the dosage parameter for each layer in the per-dosage range heads of a DRNet (- Repeat) performed worse than appending the dosage parameter on News-2, News-4 and News-8. In terms of DPE and PE, we found that DRNets outperformed all existing state-of-the-art methods with the exception of the PE on News-8, where DRNets achieved the second-best result after GPS. Lastly, the results showed that DRNet improved upon both TARNET and the MLP baseline by a large margin across all datasets - demonstrating that the hierarchical dosage subdivision introduced by DRNets is effective, and that an optimised model structure is paramount for learning representations for counterfactual inference.

Number of Dosage Strata E. To determine the impact of the choice of the number of dosage strata E on DRNet performance, we analysed the estimation performance and computation time of DRNets trained with various numbers of dosage strata E on the MVICU benchmark (Figure 5.2). With all other hyperparameters held equal, we found that a higher number of dosage strata in general improves estimation performance, because the resolution at which the dosage range is partitioned is increased. However, there is a trade-off between resolution and computational performance, as higher values of E consistently increased the computation time necessary for training and prediction.

Treatment Assignment Bias. To assess the robustness of DRNets and existing methods to increasing levels of treatment assignment bias in observational data, we compared the performance of DRNet to TARNET, MLP and GPS on the test set of News-2 with varying choices of treatment assignment bias $\kappa \in [5, 20]$ (Figure 5.3). We found that DRNet outperformed existing methods across the entire range of evaluated treatment assignment biases.

Limitations. A general limitation of methods that attempt to estimate causal effects from observational data is that they are based on untestable assumptions (Stone, 1993). In this work, we assume unconfoundedness (Imbens, 2000; Lechner, 2001), which implies that one must have reasonable certainty that the available covariate set X contains the most relevant variables for the problem setting being modelled. Making this judgement can be difficult in practice, particularly when one does not have much prior knowledge about the underlying causal process. Even without such certainty, this approach may nonetheless be a justifiable starting point to generate hypotheses when experimental data is not available (Imbens, 2004).

Table 5.2: Comparison of methods for counterfactual inference with multiple parametric treatments on News-2/4/8/16, MVICU and TCGA. We report the mean value $\pm$ the standard deviation of $\sqrt{\text{MISE}}$ on the respective test sets over 5 repeat runs with new random seeds. n.r. = not reported for computational reasons (excessive runtime). †= significantly different from DRNet ($\alpha < 0.05$).

Method	News-2	News-4	News-8	News-16	MVICU	TCGA
DRNet	8.0 ± 0.1	11.6 ± 0.1	10.2 ± 0.1	10.3 ± 0.0	31.1 ± 0.4	$\mathbf{9.6} \pm 0.0$
- Repeat	† 9.0 ± 0.1	$^{\dagger}11.9 \pm 0.2$	10.3 ± 0.1	10.4 ± 0.1	$\mathbf{31.0} \pm 0.3$	10.2 ± 0.2
+ Wasserstein	† $\mathbf{7.7} \pm 0.2$	$\mathbf{11.5} \pm 0.0$	$^{\dagger}\mathbf{10.0} \pm 0.0$	$^{\dagger}\mathbf{10.2} \pm 0.0$	32.9 ± 2.9	10.2 ± 0.9
+ PD	† 9.0 ± 0.2	$^{\dagger}12.2 \pm 0.1$	$^{\dagger}10.6 \pm 0.2$	10.3 ± 0.1	$^{\dagger}36.9 \pm 0.9$	$^{\dagger}11.9 \pm 1.4$
+ PM	† 8.4 ± 0.3	$^{\dagger}12.2 \pm 0.1$	$^{\dagger}11.4 \pm 0.3$	$^{\dagger}12.3 \pm 0.3$	31.2 ± 0.4	9.7 ± 0.2
+ PSM_{PM}	† 8.6 ± 0.1	$^{\dagger}12.2 \pm 0.2$	$^{\dagger}11.5 \pm 0.2$	$^{\dagger}12.2 \pm 0.3$	$^{\dagger}32.6 \pm 0.5$	$^{\dagger}11.4 \pm 0.6$
MLP	$^{\dagger}15.3 \pm 0.1$	$^{\dagger}14.5 \pm 0.0$	$^{\dagger}13.9 \pm 0.1$	$^{\dagger}14.0 \pm 0.0$	$^{\dagger}49.5 \pm 5.1$	$^{\dagger}15.3 \pm 0.2$
TARNET	$^{\dagger}15.5 \pm 0.1$	$^{\dagger}15.4 \pm 0.0$	$^{\dagger}14.7 \pm 0.1$	$^{\dagger}14.7 \pm 0.1$	$^{\dagger}58.0 \pm 4.8$	$^{\dagger}14.7 \pm 0.1$
GANITE	$^{\dagger}16.8 \pm 0.1$	$^{\dagger}15.6 \pm 0.1$	$^{\dagger}14.8 \pm 0.1$	$^{\dagger}14.8 \pm 0.0$	$^{\dagger}59.5 \pm 0.8$	$^{\dagger}15.4 \pm 0.2$
kNN	$^{\dagger}16.2 \pm 0.0$	$^{\dagger}14.7 \pm 0.0$	$^{\dagger}15.0 \pm 0.0$	$^{\dagger}14.5 \pm 0.0$	$^{\dagger}54.9 \pm 0.0$	n.r.
GPS	$^{\dagger}47.6 \pm 0.1$	$^{\dagger}24.7 \pm 0.1$	$^{\dagger}22.9 \pm 0.0$	$^{\dagger}15.5 \pm 0.1$	$^{\dagger}78.3 \pm 0.0$	$^{\dagger}26.3 \pm 0.0$
CF	$^{\dagger}26.0 \pm 0.0$	$^{\dagger}20.5 \pm 0.0$	$^{\dagger}19.6 \pm 0.0$	$^{\dagger}14.9 \pm 0.0$	$^{\dagger}57.5 \pm 0.0$	$^{\dagger}15.2 \pm 0.0$
BART	$^{\dagger}13.8 \pm 0.2$	$^{\dagger}14.0 \pm 0.1$	$^{\dagger}13.0 \pm 0.1$	n.r.	$^{\dagger}47.1 \pm 0.8$	n.r.

132

Table 5.3: Comparison of methods for counterfactual inference with multiple parametric treatments on News-2/4/8/16, MVICU and TCGA. We report the mean value ± the standard deviation of $\sqrt{\text{DPE}}$ on the respective test sets over 5 repeat runs with new random seeds. n.r. = not reported for computational reasons (excessive runtime). †= significantly different from DRNet ($\alpha < 0.05$).

Method	News-2	News-4	News-8	News-16	MVICU	TCGA
DRNet	14.0 ± 0.1	14.3 ± 0.2	17.2 ± 0.6	19.1 ± 2.1	**5.0** $\pm$ 1.5	2.1 ± 0.4
- Repeat	17.5 ± 7.1	†18.5 ± 3.8	16.8 ± 2.9	**16.1** ± 3.0	$5.4 \pm$ 3.1	5.4 ± 3.9
+ Wasserstein	†13.8 ± 0.1	14.2 ± 0.3	**16.7** ± 0.6	18.7 ± 2.4	20.3 ± 19.0	2.9 ± 1.3
+ PD	†**13.7** ± 0.1	14.3 ± 0.2	17.6 ± 0.0	16.5 ± 2.0	†38.9 ± 19.6	†12.7 ± 5.4
+ PM	†**13.7** ± 0.1	**13.9** ± 0.5	†21.3 ± 2.7	21.0 ± 0.7	†$12.5 \pm$ 3.5	**1.8** ± 0.1
+ PSM_{PM}	14.0 ± 0.1	14.3 ± 0.1	†19.5 ± 1.3	19.2 ± 1.6	†$28.1 \pm$ 1.4	2.5 ± 1.1
MLP	†47.1 ± 0.0	†43.2 ± 0.1	†39.8 ± 0.4	†40.7 ± 0.0	†42.8 ± 22.8	†31.1 ± 0.1
TARNET	†45.0 ± 1.1	†41.5 ± 1.1	†38.6 ± 1.1	†38.1 ± 1.4	†$106. \pm 48.8$	†38.6 ± 1.1
GANITE	†39.4 ± 0.1	†35.5 ± 0.1	†32.6 ± 0.2	†32.0 ± 0.3	†$103. \pm$ 9.1	†26.5 ± 0.6
kNN	†44.6 ± 0.0	†41.4 ± 0.0	†39.2 ± 0.0	†38.1 ± 0.0	†$60.5 \pm$ 0.0	n.r.
GPS	†42.3 ± 0.0	†34.7 ± 0.0	†22.5 ± 0.0	17.3 ± 0.1	†$81.6 \pm$ 0.0	†23.8 ± 0.0
CF	†47.1 ± 0.0	†43.2 ± 0.0	†38.3 ± 0.0	†35.2 ± 0.0	†$116. \pm$ 0.0	†30.9 ± 0.0
BART	†31.6 ± 1.4	†25.0 ± 1.5	†23.1 ± 0.4	n.r.	†$39.9 \pm$ 2.4	n.r.

Table 5.4: Comparison of methods for counterfactual inference with multiple parametric treatments on News-2/4/8/16, MVICU and TCGA. We report the mean value $\pm$ the standard deviation of $\sqrt{\text{PE}}$ on the respective test sets over 5 repeat runs with new random seeds. n.r. = not reported for computational reasons (excessive runtime). †= significantly different from DRNet ($\alpha < 0.05$).

Method	News-2	News-4	News-8	News-16	MVICU	TCGA
DRNet	15.7 ± 0.2	15.5 ± 0.4	15.1 ± 0.2	19.2 ± 14.9	12.9 ± 1.2	$\mathbf{2.1} \pm 0.1$
- Repeat	20.4 ± 9.6	$20.0 \pm 10.$	$24.3 \pm 12.$	5.3 ± 8.8	13.7 ± 2.5	2.9 ± 1.3
+ Wasserstein	$^{\dagger}15.2 \pm 0.1$	15.4 ± 0.5	15.6 ± 1.1	19.2 ± 14.9	13.6 ± 1.3	2.2 ± 0.0
+ PD	15.3 ± 0.6	$^{\dagger}32.8 \pm 0.0$	14.9 ± 0.0	† **0.9** ± 0.0	$^{\dagger}48.1 \pm 27.1$	$^{\dagger}21.8 \pm 9.2$
+ PM	$^{\dagger}\mathbf{15.1} \pm 0.1$	$\mathbf{12.4} \pm 5.3$	$^{\dagger}29.3 \pm 7.2$	33.0 ± 11.7	$\mathbf{12.1} \pm 1.9$	2.3 ± 0.4
+ PSM_{PM}	$^{\dagger}16.2 \pm 0.4$	15.3 ± 0.3	$^{\dagger}23.7 \pm 7.3$	$^{\dagger} 1.3 \pm 0.8$	$^{\dagger}23.3 \pm 4.7$	$^{\dagger}2.3 \pm 0.1$
MLP	$^{\dagger}49.5 \pm 0.1$	$^{\dagger}49.6 \pm 0.0$	$^{\dagger}48.5 \pm 0.7$	$^{\dagger}48.3 \pm 0.3$	$^{\dagger}23.7 \pm 9.8$	$^{\dagger}37.1 \pm 2.3$
TARNET	$^{\dagger}47.2 \pm 2.1$	$^{\dagger}48.0 \pm 1.8$	$^{\dagger}47.7 \pm 1.1$	$^{\dagger}44.8 \pm 3.2$	$^{\dagger}102. \pm 44.0$	$^{\dagger}47.7 \pm 1.1$
GANITE	$^{\dagger}42.6 \pm 0.3$	$^{\dagger}40.3 \pm 0.3$	$^{\dagger}42.7 \pm 0.4$	34.4 ± 0.6	$^{\dagger}96.7 \pm 9.9$	$^{\dagger}25.9 \pm 1.1$
kNN	$^{\dagger}45.2 \pm 0.0$	$^{\dagger}42.1 \pm 0.0$	$^{\dagger}45.5 \pm 0.0$	$^{\dagger}46.4 \pm 0.0$	$^{\dagger}59.1 \pm 0.0$	n.r.
GPS	$^{\dagger}44.6 \pm 0.0$	$^{\dagger}13.3 \pm 0.0$	$^{\dagger}\mathbf{13.3} \pm 0.0$	$^{\dagger}1.6 \pm 0.0$	$^{\dagger}140. \pm 0.0$	$^{\dagger}20.0 \pm 0.0$
CF	$^{\dagger}48.9 \pm 0.0$	$^{\dagger}49.6 \pm 0.0$	$^{\dagger}49.6 \pm 0.0$	$^{\dagger}48.3 \pm 0.0$	$^{\dagger}108. \pm 0.0$	$^{\dagger}35.3 \pm 0.0$
BART	$^{\dagger}35.5 \pm 14.$	$^{\dagger}34.6 \pm 4.2$	$^{\dagger}44.5 \pm 1.2$	n.r.	13.2 ± 1.0	n.r.

134

5.6 Conclusion

We presented a deep-learning approach to learning to estimate individual dose-response to multiple treatments with continuous dosage parameters based on observational data. We extended several existing regularisation strategies to the setting with any number of treatment options with associated dosage parameters, and combined them with our approach in order to address treatment assignment bias inherent in observational data. In addition, we introduced performance metrics, model selection criteria, model architectures, and new open benchmarks for this setting. Our experiments demonstrated that model structure is paramount in learning neural representations for counterfactual inference of dose-response curves from observational data, and that there is a trade-off between model resolution and computational performance in DRNets. DRNets significantly outperform existing state-of-the-art methods in inferring individual dose-response curves across several benchmarks.

5.7 Supplementary Material

5.7.A Source Code

The source code for this work is available at:
`https://github.com/d909b/drnet`.

5.7.B Treatment Assignment Bias Regularisation

To address treatment assignment bias in DRNets, we evaluated four different regularisation strategies: (1) distribution matching between treatment groups using the Wasserstein regulariser (+ Wasserstein), (2) propensity dropout (+ PD), (3) matching on the entire dataset, and (4) matching on the batch level. Because neither of these regularisation strategies were originally developed for the setting with parametric

treatment options, we implemented näive extensions thereof for this setting. To extend (1) to this setting, we followed Schwab et al. (2018b) and penalised pair-wise differences between treatment group distributions in the topmost shared hidden layer using the first treatment option as the control treatment. For (2), we applied PD for each treatment option to both the corresponding per-treatment layers and to the respective treatment option's head layers in each dosage stratum. To use (3) dataset-wide and (4) per-batch matching with parametric treatments, we followed the PM algorithm outlined in (Schwab et al., 2018b), and matched directly on the covariates X with the dosage parameter s added to the covariate set. For datasets of dimensionality higher than 200, we matched on a low-dimensional representation obtained via Principal Components Analysis (PCA) using 50 principal components in order to reduce the computational requirements.

5.7.C Hyperparameters

We used a standardised approach to hyperparameter optimisation for all methods. Each method was given exactly the same amount of hyperparameter optimisation runs (5 on TCGA, and 10 on all other benchmarks). We also fixed the random seed such that all methods were evaluated on exactly the same random hyperparameter configurations. For the methods based on neural network models (DRNet, + Repeat, + Wasserstein, + PD, + PM, + PSM_{PM}, MLP, TARNET, GANITE), we chose hyperparameters at random from predefined ranges (Table 5.5). For "+ Wasserstein", we additionally chose an imbalance penalty weight at random from 0.1, 1.0, or 10.0. For GANITE, we also randomly chose the supervised loss weights α and β from 0.1, 1.0, and 10.0 (Yoon et al., 2018)) as an additional hyperparameter. For BART and CF, we used the default hyperparameters from their respective implementations in the R-packages "bartMachine" (Kapelner & Bleich, 2016) and "grf" (Athey et al., 2019). Because CF was designed for estimating the difference in treatment effect between two treatment options and not for directly estimating treatment

Table 5.5: Hyperparameter ranges used in our experiments.

Hyperparameter	Values
Batch size B	32, 64, 128
Number of units per hidden layer M	24, 48, 96
Number of hidden layers L	2, 3
Dropout percentage p_{dropout}	$[0.0, 0.2]$

outcomes $\hat{y}_t$, we used a baseline ridge regression model with regularisation strength $\alpha = 0.5$ to estimate a control outcome $\hat{y}_0$ for the first treatment and one CF model to estimate the difference in treatment effect between that control treatment and all other treatment options (Schwab et al., 2018b). For kNN, we used 5 nearest neighbours to compute the potential outcomes matching on the covariates X with the dosage parameter s added as an additional covariate. For GPS, we used the implementation in the "causaldrf" R-package (Galagate, 2016) with a normal treatment model, a linear treatment formula, and a polynomial of degree 2 as the outcome formula. To reduce the computational requirements for GPS to a manageable level, we additionally preprocessed the covariates X using PCA dimensionality reduction with 16 principal components for benchmarks with a covariate-space dimensionality higher than 200.

This page was intentionally left blank.

Chapter 6

Conclusion

"An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question." - John Tukey

The main goal of this work was to advance the state-of-the-art in the use of deep learning for medical applications. Building on the deep learning framework, we developed several approaches that enable the use of deep neural networks and large-scale health data to treat, explain and diagnose. We evaluated our proposed approaches using data from several challenging real-world tasks from the medical domain. Our experimental results demonstrated that our proposed approaches advance the state-of-the-art in deep learning for medical applications along multiple dimensions, including in learning with fewer expert labels when large amounts of unlabelled time series data are available, diagnosing based on information from multiple heterogenous data streams, learning to explain decisions of neural networks, and estimating individual treatment effects in the setting with multiple available treatment options with associated continuous dosage parameters. Finally, we showcased the potential of using deep learning and large-scale health data to improve healthcare by developing and evaluating machine-learning systems for reducing false alarms in critical care, diagnosing Parkinson's disease from smartphone data, identifying discriminatory genes across cancer types, and estimating optimal treatment and dosage choices for mechanical ventilation in critical care and cancer treatment.

6.1 Summary and Discussion

In the following paragraphs, we briefly summarise the individual results presented in this work and discuss their limitations and implications.

Distantly Supervised Multitask Learning in Critical Care. In Chapter 2, we introduced a deep-learning approach to reducing the false alarm rate in critical care. In this setting, as is the case in many medical applications, large amounts of unlabelled data are readily available from patient monitoring systems, but validated labels from medical experts are expensive and time-consuming to collect. We therefore developed a semi-supervised approach that was based on the idea of automatically selecting multiple auxiliary tasks for distant multitask supervision. As candidate auxiliary tasks, we leveraged a large repository of manually-engineered time series features. We then introduced a task selection strategy based on task importance relative to the main task to select a subset of the candidate for distant multitask supervision. We used the selected auxiliary tasks, a specifically engineered neural network architecture, and a specialised learning procedure in order to learn to assess alarms with fewer available expert labels. To evaluate our approach, we used a dataset consisting of almost 14 000 alarms collected by our clinical collaborators in the Neurocritical Care Unit of the University of Zurich, Switzerland. Our results indicated that intelligent false alarm reduction systems could in the future potentially play a role in addressing the problem of alarm desensitisation in critical care, and that distant multitask supervision with task-specific auxiliary tasks may in certain settings outperform existing state-of-the-art methods for semi-supervised learning that rely on task-unspecific auxiliary tasks, such as reconstruction objectives.

While our experimental results are encouraging, prospective experiments are needed to determine conclusively whether false alarm reduction algorithms can improve outcomes (i) in terms of working conditions for clinical staff and (ii) in terms of overall risk for patients. In addition, our experimental evaluation only covered a limited set of patients, biosignal

types, monitoring system configurations, monitoring modalities (Muroi et al., 2019), and monitoring days, and it is therefore not yet known whether our results generalise to all possible monitoring configurations, patient types, biosignals, and alarm-generating algorithms that may be encountered in the wild. As there may, for example, be rare types of alarms and patient conditions that were not observed in our dataset. More data from a larger number of patients and monitoring system configurations is needed to evaluate to what degree the performance of our proposed false alarm reduction approach is influenced by environmental, patient-specific, and configuration-specific factors. Finally, non-technical questions, such as how to best integrate the proposed false alarm reduction system into clinical care and how to address regulatory issues associated with lowering the degree of urgency with which alarms are reported, need to be answered before such a system could be used in practice.

Learning to Diagnose Parkinson's Disease from Smartphone Data. In Chapter 3, we presented a hierarchical deep-learning approach to learning to diagnose Parkinson's disease based on multiple, heterogenous data streams collected over long periods of time. Our approach, which we called evidence aggregation model (EAM), built on the idea of separating learning to assess individual data streams, and learning to integrate information from multiple data streams over time into two distinct stages. In addition, we extended our deep-learning approach with a hierarchical neural soft attention mechanism in order to quantify the importance of individual tests and segments within those tests towards the model's final diagnostic prediction. To evaluate our proposed approach, we performed an experimental evaluation on the mPower cohort consisting of 1 853 people - the largest cohort used for validating a machine-learning algorithm for diagnosing PD based on smartphone data to date (Schwab & Karlen, 2019b). Participants in the mPower cohort contributed their demographic profiles, prior professional diagnoses, and smartphone monitoring data from several types of tests over an average of around 25 days. We also qualitatively analysed which segments within individual tests were considered important by the

model by inspecting the outputs of the hierarchical neural soft attention mechanism. Our results indicated that deep learning and smartphone monitoring data collected over long periods of time could in the future be used as a digital biomarker for the diagnosis of Parkinson's disease (Schwab & Karlen, 2019b), and that the presented hierarchical approach to learning to diagnose based on multiple heterogenous data streams may in certain settings outperform end-to-end learning and aggregation baselines.

In addition to diagnostic decision support, a potentially interesting use case of smartphone-derived digital biomarkers for Parkinson's disease would be to screen for people in the early stages of Parkinson's disease that do not yet have a professional diagnosis. However, our dataset offered limited insights into this group, since the cohort we used was on average around 60 years old, and because we did not have data on whether any of the younger patients would potentially have received a Parkinson's diagnosis in the future. Because only few would go on to develop Parkinson's disease, a larger group of people would have to be recruited and followed up with for longer periods of time in order to collect a representative amount of smartphone monitoring data on early-stage Parkinson's disease. In future work, it would be interesting to extend the presented approach to diagnosing Parkinson's disease to passively-collected smartphone monitoring data, since maintaining user engagement over long periods of time with a small number of user-initiated tests can be difficult. In addition, it is likely that similar approaches could successfully be applied to other disorders with motor, voice, cognitive, or dexterity symptoms, and to other tasks, such as symptom monitoring (Zhan et al., 2018) and predicting disease progression (Küffner et al., 2015).

Learning Important Features with Neural Networks. In Chapter 4, we described an approach to learning to jointly produce accurate estimates of feature importance and predictions in a single neural network model. Our approach utilised an attentive mixture of experts (AME) structure that uses attentive gating networks to control the contributions of individual experts to the final model output. To ensure the assigned attention factors

accurately reflect feature importance, we introduced a secondary Granger-causal objective. We performed extensive experiments on three datasets with various properties that demonstrated that AMEs produce estimates of feature importance that compare favourably to those produced by existing state-of-the-art methods, that AMEs are significantly faster than existing methods at producing those estimates, and that the associations reported by AMEs are consistent with those reported by domain experts (Schwab et al., 2019b). Our results confirmed that the proposed Granger-causal objective can be used to train neural networks to learn to accurately estimate feature importance alongside its predictions. Neural network models that can produce estimates of feature importance and predictions are particularly useful in medical applications where the requirements towards interpretability are high. In addition, AMEs may be a good alternative to existing feature importance estimation methods for neural networks in settings with large numbers of input features, where AMEs are significantly faster at producing feature importance estimates than existing methods. Interesting questions for future research would be how this approach could be used to train explanation models for pre-trained models with any model architecture (Schwab & Karlen, 2019a), and how this approach could be extended to produce other types of explanations, such as concept-based (Kim et al., 2018) or textual (Zhang et al., 2017) explanations.

Learning to Estimate Individual Dose Response. In Chapter 5, we introduced a deep-learning approach to learning to estimate individual treatment effects from observational data in the setting with multiple available treatment options with associated continuous dosage parameters. Our approach was based on a hierarchical dose response network (DRNet) architecture that ensures the importance of both the treatment indicator as well as the continuous dosage parameter is maintained. We additionally extended our approach with several existing regularisation strategies that address treatment assignment bias in observational data. To compare our proposed approach to existing state-of-the-art methods for learning to estimate individual treatment effects, we developed three performance metrics and three benchmarks with different properties for this setting,

including a benchmark on treatment recommendation for mechanical ventilation in critical care and a benchmark on cancer treatment. We found that DRNets set a new state-of-the-art in learning to estimate individual treatment effects from observational data in the setting with multiple available treatment options with associated continuous dosage parameters across all the evaluated benchmark datasets and benchmark configurations. Models that estimate individual treatment effects from observational data could, under certain assumptions, in the future potentially be used to provide treatment recommendations in medical settings.

One of the main limitations of learning to estimate counterfactual outcomes from observational data is that we only ever observe the realised outcome, and never any of the other possible counterfactual outcomes. The inability to observe counterfactual outcomes makes the evaluation of estimators for counterfactual outcomes difficult in settings in which we do not have access to the exact outcome-generating process. This is a considerable problem in medical applications, where we would ideally like to have a priori estimates of how well our treatment recommendation model would perform at its intended task. To facilitate the application of estimators of individual treatment effects for treatment recommendation in medicine, it is therefore imperative that better tools for evaluating their generalisation capabilities are developed in the future. Interesting avenues to explore in the future for this purpose would be the combination of observational data with limited amounts of experimental data (Kallus et al., 2018), information-theoretic model validation (Buhmann, 2010), or custom cross-validation methods (Athey & Imbens, 2016).

A limitation of all the works enclosed in this thesis is that we have, due to the effort associated with data collection and implementation, only evaluated each of the methods on a limited number of different tasks from the medical domain. It is thus difficult to estimate to what degree our results generalise to other tasks, other diseases, and other data modalities. Further research on a broader array of medical tasks is necessary to determine under what conditions the presented methods generalise to other settings.

6.2 Outlook

We expect that the adoption of machine learning for medical applications will continue to expand in the future. As demonstrated by this and related works (Ganna & Ingelsson, 2015; Doherty et al., 2017; Alaa et al., 2019), large-scale health databases contain vast amounts of information about personal health and well-being that could in the future potentially be used to improve decision-making in healthcare. The advantages of using machine learning for healthcare applications are manifold: Firstly, algorithms trained on representative datasets could, once validated, be made available for deployment across many institutions, hospitals, and geographies to ensure that the best evidence-based care is available to a large number of people. Secondly, widespread adoption of predictive models could potentially improve decision-making in healthcare by integrating the vast amounts of personal health data available today. At present, many informative data sources, such as data from wearables and smart devices, are not well integrated into clinical decision-making due to concerns about data sharing and privacy (Lo, 2015). Increasing adoption of medical decision-support systems based on machine-learning methods could provide an impetus for better data sharing and integration practices in medicine, and privacy-preserving generative models may facilitate sharing for datasets where legal or privacy concerns would otherwise prevent data sharing (Esteban et al., 2017; Beaulieu-Jones et al., 2019). Thirdly, predictive models could potentially in some use cases be used on-demand and without the need for expensive and time-consuming in-person visits, which could potentially allow expanding care to underserved geographic regions. Fourthly, the use of predictive models for time-consuming tasks could potentially free up time for physicians to focus on patients' individual needs, and improve the overall working environment for healthcare workers. Lastly, predictive models are, once trained and integrated into clinical processes, inexpensive to distribute and maintain, and could therefore potentially over time achieve considerable cost savings for healthcare systems. While the potential of machine learning in medical applications is high, there are also risks

associated with its use. Regulators, healthcare providers, and data owners should therefore carefully weigh both benefits and risks, and work on ensuring the secure and ethical use of healthcare data for good (Horvitz & Mulligan, 2015). Overall, we are enthusiastic about the potential future of machine learning in healthcare applications, and believe that it will, ultimately, play an important role in improving human health and wellbeing in the future.

"The best way to predict the future is to create it." - Abraham Lincoln

Appendix

This page was intentionally left blank.

Appendix A

Curriculum Vitae

Patrick Schwab

Education

Doctor of Science, ETH Zürich, Switzerland	2017 - 2019
Master of Science, University of Vienna, Austria	2013 - 2015
Bachelor of Science, Technikum Vienna, Austria	2010 - 2013

Publications

Schwab, Patrick, and Karlen, Walter. CXPlain: Causal explanations for model interpretation under uncertainty. *Advances in Neural Information Processing Systems (to appear)*, 2019

Schwab, Patrick, and Karlen, Walter. A deep learning approach to diagnosing multiple sclerosis from smartphone data. *(in submission)*, 2019

Muroi, Carl, Meier, Sandro, De Luca, Valeria, Mack, David J., Strässle, Christian, **Schwab, Patrick**, Karlen, Walter and Keller, Emanuela. Automated false alarm reduction in a real-life intensive care setting using motion detection. *Neurocritical Care*, 2019

Schwab, Patrick, Linhardt, Lorenz, Bauer, Stefan, Buhmann, Joachim. M., and Karlen, Walter. Learning counterfactual representations for estimating individual dose-response curves. *arXiv preprint arXiv:1902.00981 (in submission)*, 2019

Schwab, Patrick and Karlen, Walter. PhoneMD: Learning to diagnose Parkinson's disease from smartphone data. *AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, United States, 27 Jan - 1 Feb, 2019

Schwab, Patrick, Miladinovic, Djordje and Karlen, Walter. Granger-causal attentive mixtures of experts: Learning important features with neural networks. *AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, United States, 27 Jan - 1 Feb, 2019

Schwab, Patrick, Linhardt, Lorenz and Karlen, Walter. Perfect Match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656 (in submission)*, 2018

Schwab, Patrick, Keller, Emanuela, Muroi, Carl, Mack, David J., Strässle, Christian and Karlen, Walter. Not to cry wolf: Distantly supervised multitask learning in critical care. *International Conference on Machine Learning*, Stockholm, Sweden, 10-15 July, 2018

Schwab, Patrick, Scebba, Gaetano C., Zhang, Jia, Delai, Marco and Karlen, Walter. Beat by beat: Classifying cardiac arrhythmias with recurrent neural networks. *Computing in Cardiology*, Rennes, France, Sept 24-27, 2017

Schwab, Patrick, and Helmut Hlavacs. Capturing the essence: Towards the automated generation of transparent behavior models. *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Santa Cruz, California, United States, November 14-18, 2015

Schwab, Patrick, and Helmut Hlavacs. PALAIS: A 3D simulation environment for artificial intelligence in games. *AISB Convention*, Canterbury, United Kingdom, April 20-22, 2015

Zaharieva, Maia, and **Patrick Schwab**. A unified framework for retrieving diverse social images. *MediaEval 2014 Workshop*, Barcelona, Catalunya, Spain, October 16-17, 2014

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Aboukhalil, A., Nielsen, L., Saeed, M., Mark, R. G., and Clifford, G. D. Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. *Journal of Biomedical Informatics*, 41(3):442–451, 2008.

Aburatani, H., Ishikawa, S., and Nakano, K. Diagnosis and treatment of cancer using anti-TMPRSS11E antibody, July 14 2015. US Patent 9,079,957.

Adler, P., Falk, C., Friedler, S. A., Rybeck, G., Scheidegger, C., Smith, B., and Venkatasubramanian, S. Auditing black-box models for indirect influence. In *International Conference on Data Mining*, pp. 1–10, 2016.

Adler-Milstein, J., DesRoches, C. M., Kralovec, P., Foster, G., Worzala, C., Charles, D., Searcy, T., and Jha, A. K. Electronic health record adoption in US hospitals: Progress continues, but challenges persist. *Health Affairs*, 34 (12):2174–2180, 2015.

Alaa, A. and Schaar, M. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pp. 129–138, 2018.

Alaa, A. M. and van der Schaar, M. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pp. 3424–3432, 2017.

Alaa, A. M., Weisz, M., and van der Schaar, M. Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966*, 2017.

Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. F., and van der Schaar, M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS ONE*, 14(5):1–17, 05 2019.

Ando, R. K. and Zhang, T. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

Andrews, R., Diederich, J., and Tickle, A. B. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-based Systems*, 8(6):373–389, 1995.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

Arora, S., Venkataraman, V., Zhan, A., Donohue, S., Biglan, K., Dorsey, E., and Little, M. Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. *Parkinsonism & Related Disorders*, 21(6):650–653, 2015.

Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Athey, S., Tibshirani, J., and Wager, S. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Aytar, Y., Pfaff, T., Budden, D., Paine, T., Wang, Z., and de Freitas, N. Playing hard exploration games by watching YouTube. In *Advances in Neural Information Processing Systems*, pp. 2930–2941, 2018.

Ba, J., Mnih, V., and Kavukcuoglu, K. Multiple object recognition with visual attention. In *International Conference on Learning Representations*, 2014.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

Ballinger, B., Hsieh, J., Singh, A., Sohoni, N., Wang, J., Tison, G. H., Marcus, G. M., Sanchez, J. M., Maguire, C., Olgin, J. E., and Pletcher, M. J. DeepHeart: Semi-Supervised sequence learning for cardiovascular risk prediction. In *AAAI Conference on Artificial Intelligence*, 2018.

Baxter, J. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12(149-198):3, 2000.

Beam, A. L. and Kohane, I. S. Big data and machine learning in health care. *JAMA*, 319(13):1317–1318, 2018.

Beaulieu-Jones, B. K., Wu, Z. S., Williams, C., Lee, R., Bhavnani, S. P., Byrd, J. B., and Greene, C. S. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, 2019.

Ben-David, S. and Schuller, R. Exploiting task relatedness for multiple task learning. *Lecture Notes in Computer Science*, pp. 567–580, 2003.

Bengio, E., Bacon, P.-L., Pineau, J., and Precup, D. Conditional computation in neural networks for faster models. *arXiv preprint arXiv:1511.06297*, 2015.

Blaschko, M., Vedaldi, A., and Zisserman, A. Simultaneous object detection and ranking with weak supervision. In *Advances in Neural Information Processing Systems*, pp. 235–243, 2010.

Bot, B. M., Suver, C., Neto, E. C., Kellen, M., Klein, A., Bare, C., Doerr, M., Pratap, A., Wilbanks, J., Dorsey, E. R., et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. *Scientific Data*, 3:160011, 2016.

Bothwell, L. E., Greene, J. A., Podolsky, S. H., and Jones, D. S. Assessing the Gold Standard — lessons from the history of RCTs. *New England Journal of Medicine*, 374(22):2175–2181, 2016.

Breiman, L. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Brenner, D. R., Brennan, P., Boffetta, P., Amos, C. I., Spitz, M. R., Chen, C., Goodman, G., Heinrich, J., Bickeböller, H., Rosenberger, A., et al. Hierarchical modeling identifies novel lung cancer susceptibility variants in inflammation pathways among 10,140 cases and 11,012 controls. *Human Genetics*, 132(5):579–589, 2013.

Buhmann, J. M. Information theoretic model validation for clustering. In *IEEE International Symposium on Information Theory*, pp. 1398–1402, 2010.

Campbell, C. L., Smyth, S., Montalescot, G., and Steinhubl, S. R. Aspirin dose for the prevention of cardiovascular disease: A systematic review. *JAMA*, 297(18):2018–2024, 2007.

Cao, Q., Li, Y.-Y., He, W.-F., Zhang, Z.-Z., Zhou, Q., Liu, X., Shen, Y., and Huang, T.-T. Interplay between microRNAs and the STAT3 signaling pathway in human cancers. *Physiological Genomics*, 45(24):1206–1214, 2013.

Carpenter, D. *Reputation and power: Organizational image and pharmaceutical regulation at the FDA*. Princeton University Press, 2014.

Castelvecchi, D. Can we open the black box of AI? *Nature News*, 538(7623): 20, 2016.

Char, D. S., Shah, N. H., and Magnus, D. Implementing machine learning in health care — addressing ethical challenges. *The New England Journal of Medicine*, 378(11):981, 2018.

Che, Z., Purushotham, S., Khemani, R., and Liu, Y. Interpretable deep models for ICU outcome prediction. In *AMIA Annual Symposium Proceedings*, volume 2016, pp. 371. American Medical Informatics Association, 2016.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.

Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., and Zhang, Z. MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015.

Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., and Li, L. China Kadoorie Biobank of 0.5 million people: Survey methods, baseline characteristics and long-term follow-up. *International Journal of Epidemiology*, 40(6):1652–1666, 2011.

Cheng, L.-F., Darnell, G., Chivers, C., Draugelis, M. E., Li, K., and Engelhardt, B. E. Sparse multi-output Gaussian processes for medical time series prediction. *arXiv preprint arXiv:1703.09112*, 2017.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387, 2018.

Chipman, H. and McCulloch, R. BayesTree: Bayesian additive regression trees. *R package version 0.3-1.4*, 2016.

Chipman, H. A., George, E. I., and McCulloch, R. E. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. Doctor AI: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pp. 301–318, 2016a.

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., and Stewart, W. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pp. 3504–3512, 2016b.

Chollet, F. et al. Keras, 2015. URL `https://github.com/fchollet/keras`.

Christ, M., Kempa-Liehr, A. W., and Feindt, M. Distributed and parallel time series feature extraction for industrial big data applications. *arXiv preprint arXiv:1610.07717*, 2016.

Clifford, G., Liu, C., Moody, B., Lehman, L., Silva, I., Li, Q., Johnson, A., and Mark, R. G. AF classification from a short single lead ECG recording: The Physionet Computing in Cardiology Challenge 2017. *Computing in Cardiology*, 44, 2017.

Clifford, G. D., Silva, I., Moody, B., Li, Q., Kella, D., Shahin, A., Kooistra, T., Perry, D., and Mark, R. G. The PhysioNet/Computing in Cardiology Challenge 2015: Reducing false arrhythmia alarms in the ICU. In *Computing in Cardiology*, pp. 273–276. IEEE, 2015.

Clifton, L., Clifton, D. A., Pimentel, M. A., Watkinson, P. J., and Tarassenko, L. Gaussian processes for personalized e-health monitoring with wearable sensors. *IEEE Transactions on Biomedical Engineering*, 60(1):193–197, 2012.

Clifton, L., Clifton, D. A., Pimentel, M. A., Watkinson, P. J., and Tarassenko, L. Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors. *IEEE Journal of Biomedical and Health Informatics*, 18(3):722–730, 2013.

Collins, F. S. and Varmus, H. A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795, 2015.

Connolly, B. S. and Lang, A. E. Pharmacological treatment of Parkinson disease: A review. *JAMA*, 311(16):1670–1683, 2014.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

Cvach, M. Monitor alarm fatigue: An integrative review. *Biomedical Instrumentation & Technology*, 46(4):268–277, 2012.

Dai, Z., Yang, Z., Yang, F., Cohen, W. W., and Salakhutdinov, R. Good semi-supervised learning that requires a bad GAN. In *Advances in Neural Information Processing Systems*, 2017.

Darcy, A. M., Louie, A. K., and Roberts, L. W. Machine learning and the profession of medicine. *JAMA*, 315(6):551–552, 2016.

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342, 2018.

De Lau, L. M. and Breteler, M. M. Epidemiology of Parkinson's disease. *The Lancet Neurology*, 5(6):525–535, 2006.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Deriu, J., Lucchi, A., De Luca, V., Severyn, A., Müller, S., Cieliebak, M., Hofmann, T., and Jaggi, M. Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In *International Conference on World Wide Web*, pp. 1045–1052, 2017.

Doersch, C. and Zisserman, A. Multi-task self-supervised visual learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2051–2060, 2017.

Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., White, T., Van Hees, V. T., Trenell, M. I., Owen, C. G., et al. Large scale population assessment of physical activity using wrist worn accelerometers: The UK Biobank study. *PloS ONE*, 12(2):e0169649, 2017.

Dorsey, E. R., Papapetropoulos, S., Xiong, M., and Kieburtz, K. The first frontier: Digital biomarkers for neurodegenerative disorders. *Digital Biomarkers*, 1(1):6–13, 2017.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Drew, B. J., Harris, P., Zègre-Hemsey, J. K., Mammone, T., Schindler, D., Salas-Boni, R., Bai, Y., Tinoco, A., Ding, Q., and Hu, X. Insights into the problem of alarm fatigue with physiologic monitor devices: A comprehensive observational study of consecutive intensive care unit patients. *PLoS ONE*, 9(10):e110274, 2014.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Dzau, V. J. and Balatbat, C. A. Health and societal implications of medical and technological advances. *Science Translational Medicine*, 10 (463):eaau4778, 2018.

Eerikäinen, L. M., Vanschoren, J., Rooijakkers, M. J., Vullings, R., and Aarts, R. M. Decreasing the false alarm rate of arrhythmias in intensive care using a machine learning approach. In *Computing in Cardiology*, 2015.

Emrani, S., McGuirk, A., and Xiao, W. Prognosis and Diagnosis of Parkinson's Disease Using Multi-Task Learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1457–1466. ACM, 2017.

Esteban, C., Hyland, S. L., and Rätsch, G. Real-valued (medical) time series generation with recurrent conditional GANs. *arXiv preprint arXiv:1706.02633*, 2017.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

Fagerberg, L., Hallström, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & Cellular Proteomics*, 13(2):397–406, 2014.

Fallet, S., Yazdani, S., and Vesin, J.-M. A multimodal approach to reduce false arrhythmia alarms in the intensive care unit. In *Computing in Cardiology*, 2015.

Ferguson, N. D., Fan, E., Camporota, L., Antonelli, M., Anzueto, A., Beale, R., Brochard, L., Brower, R., Esteban, A., Gattinoni, L., et al. The Berlin definition of ARDS: An expanded rationale, justification, and supplementary material. *Intensive Care Medicine*, 38(10):1573–1582, 2012.

Fernando, B., Bilen, H., Gavves, E., and Gould, S. Self-supervised video representation learning with odd-one-out networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5729–5738, 2017.

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., and Kohane, I. S. Adversarial attacks on medical machine learning. *Science*, 363(6433): 1287–1289, 2019.

Flores, B. E. A pragmatic view of accuracy measurement in forecasting. *Omega*, 14(2):93–98, 1986.

Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision*, 2017.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767, 2011.

Galagate, D. *Causal inference with a continuous treatment and outcome: Alternative estimators for parametric dose-response functions with applications.* PhD thesis, University of Maryland, 2016.

Ganna, A. and Ingelsson, E. 5 year mortality predictors in 498 103 UK Biobank participants: A prospective population-based study. *The Lancet*, 386(9993):533–540, 2015.

Gaziano, J. M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology*, 70:214–223, 2016.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010.

Gervasini, G., Carrillo, J. A., Garcia, M., San Jose, C., Cabanillas, A., and Benitez, J. Adenosine triphosphate-binding cassette B1 (ABCB1)(multidrug resistance 1) G2677T/A gene polymorphism is associated with high risk of lung cancer. *Cancer*, 107(12):2850–2857, 2006.

Ghassemi, M., Naumann, T., Doshi-Velez, F., Brimmer, N., Joshi, R., Rumshisky, A., and Szolovits, P. Unfolding physiological state: Mortality modelling in intensive care units. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 75–84. ACM, 2014.

Ghassemi, M., Pimentel, M. A., Naumann, T., Brennan, T., Clifton, D. A., Szolovits, P., and Feng, M. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In *AAAI Conference on Artificial Intelligence*, pp. 446–453, 2015.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., and Ranganath, R. Opportunities in Machine Learning for Healthcare. *arXiv preprint arXiv:1806.00388*, 2018.

Global Parkinson's Disease Survey Steering Committee. Factors impacting on quality of life in Parkinson's disease: Results from an international survey. *Movement Disorders*, 17(1):60–67, 2002.

Goetz, C. G., Poewe, W., Rascol, O., and Sampaio, C. Evidence-based medical review update: Pharmacological and surgical treatments of Parkinson's disease: 2001 to 2004. *Movement Disorders*, 20(5):523–539, 2005.

Goldberg, Y. and Levy, O. word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

Gong, J.-P., Yang, L., Tang, J.-W., Sun, P., Hu, Q., Qin, J.-W., Xu, X.-M., Sun, B.-C., and Tang, J.-H. Overexpression of microRNA-24 increases the sensitivity to paclitaxel in drug-resistant breast carcinoma cell lines via targeting ABCB9. *Oncology Letters*, 12(5):3905–3911, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT Press, 2016.

Goodman, B. and Flaxman, S. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017.

Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

Graves, A. and Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

Graves, A., Jaitly, N., and Mohamed, A.-r. Hybrid speech recognition with deep bidirectional LSTM. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 273–278. IEEE, 2013.

Grnarova, P., Schmidt, F., Hyland, S. L., and Eickhoff, C. Neural document embeddings for intensive care patient mortality prediction. *arXiv preprint arXiv:1612.00467*, 2016.

Hammerla, N. Y., Fisher, J., Andras, P., Rochester, L., Walker, R., and Plötz, T. PD Disease State Assessment in Naturalistic Environments Using Deep Learning. In *AAAI Conference on Artificial Intelligence*, 2015.

Han, J.-Y., Lim, H.-S., Yoo, Y.-K., Shin, E. S., Park, Y. H., Lee, S. Y., Lee, J.-E., Lee, D. H., Kim, H. T., and Lee, J. S. Associations of ABCB1, ABCC2, and ABCG2 polymorphisms with irinotecan-pharmacokinetics and clinical outcome in patients with advanced non-small cell lung cancer. *Cancer*, 110(1):138–147, 2007.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Hernán, M. A. and Robins, J. M. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764, 2016.

Hlaváč, V., Brynychová, V., Václavíková, R., Ehrlichová, M., Vrána, D., Pecha, V., Koževnikovová, R., Trnková, M., Gatěk, J., Kopperová, D., et al. The expression profile of ATP-binding cassette transporter genes in breast carcinoma. *Pharmacogenomics*, 14(5):515–529, 2013.

Hlavata, I., Mohelnikova-Duchonova, B., Vaclavikova, R., Liska, V., Pitule, P., Novak, P., Bruha, J., Vycital, O., Holubec, L., Treska, V., et al. The role of ABC transporters in progression and clinical outcome of colorectal cancer. *Mutagenesis*, 27(2):187–196, 2012.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236, 2007.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011.

Honan, L., Funk, M., Maynard, M., Fahs, D., Clark, J. T., and David, Y. Nurses' perspectives on clinical alarms. *American Journal of Critical Care*, 24(5):387–395, 2015.

Horvitz, E. and Mulligan, D. Data, privacy, and the greater good. *Science*, 349(6245):253–255, 2015.

Hyndman, R. J. and Khandakar, Y. *Automatic time series for forecasting: The forecast package for R*. Monash University, Department of Econometrics and Business Statistics, 2007.

Ilse, M., Tomczak, J. M., and Welling, M. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018.

Imbens, G. W. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000.

Imbens, G. W. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456, 2015.

Itti, L. and Baldi, P. F. Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems*, pp. 547–554, 2006.

Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., and Kavukcuoglu, K. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2017.

Jansová, E., Koutna, I., Krontorad, P., Svoboda, Z., Křivánková, S., Žaloudík, J., Kozubek, M., and Kozubek, S. Comparative transcriptome maps: A new approach to the diagnosis of colorectal carcinoma patients using cDNA microarrays. *Clinical Genetics*, 69(3):218–227, 2006.

Jensen, P. B., Jensen, L. J., and Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395, 2012.

Jiang, Y., Huang, Y., Du, Y., Zhao, Y., Ren, J., Ma, S., and Wu, C. Identification of prognostic genes and pathways in lung adenocarcinoma using a bayesian approach. *Cancer Informatics*, 1:7, 2017.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pp. 3020–3029, 2016.

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

Jordan, M. I. and Jacobs, R. A. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

Jordan, M. I. and Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

Kallus, N. Recursive partitioning for personalization using observational data. In *International Conference on Machine Learning*, 2017.

Kallus, N., Puli, A. M., and Shalit, U. Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems*, pp. 10911–10920, 2018.

Kapelner, A. and Bleich, J. bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40, 2016.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2017.

Kendall, M. G. The treatment of ties in ranking problems. *Biometrika*, pp. 239–251, 1945.

Kim, B., M., W., Gilmer, J., C., C., J., W., , Viegas, F., and Sayres, R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors. In *International Conference on Machine Learning*, 2018.

Kim, G. R., Ku, Y. J., Cho, S. G., Kim, S. J., and Min, B. S. Associations between gene expression profiles of invasive breast cancer and breast imaging reporting and data system MRI lexicon. *Annals of Surgical Treatment and Research*, 93(1):18–26, 2017.

Kimura, S., Sato, T., Ikeda, S., Noda, M., and Nakayama, T. Development of a database of health insurance claims: Standardization of disease classifications and anonymous record linkage. *Journal of Epidemiology*, 20 (5):413–419, 2010.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un) reliability of saliency methods. *arXiv preprint arXiv:1711.00867*, 2017.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pp. 3581–3589, 2014.

Klucken, J., Barth, J., Kugler, P., Schlachetzki, J., Henze, T., Marxreiter, F., Kohl, Z., Steidl, R., Hornegger, J., Eskofier, B., and Winkler, J. Unbiased and mobile gait analysis detects motor impairment in Parkinson's disease. *PLoS ONE*, 8(2):e56956, 2013.

Knaus, M. C., Lechner, M., and Strittmatter, A. Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *arXiv preprint arXiv:1810.13237*, 2018.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. *International Conference of Machine Learning*, 2017.

Kooi, T., Litjens, G., Van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., and Karssemeijer, N. Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, 35:303–312, 2017.

Krasteva, V., Jekova, I., Leber, R., Schmid, R., and Abächerli, R. Real-time arrhythmia detection with supplementary ECG quality and pulse wave monitoring for the reduction of false alarms in ICUs. *Physiological Measurement*, 37(8):1273, 2016.

Krause, J., Perer, A., and Ng, K. Interacting with predictions: Visual inspection of black-box machine learning models. In *CHI Conference on Human Factors in Computing Systems*, pp. 5686–5697. ACM, 2016.

Kringen, M. K., Stormo, C., Grimholt, R. M., Berg, J. P., and Piehler, A. P. Copy number variations of the ATP-binding cassette transporter ABCC6 gene and its pseudogenes. *BMC Research Notes*, 5(1):425, 2012.

Krøigård, A. B., Larsen, M. J., Lænkholm, A.-V., Knoop, A. S., Jensen, J. D., Bak, M., Mollenhauer, J., Thomassen, M., and Kruse, T. A. Identification of metastasis driver genes by massive parallel sequencing of successive steps of breast cancer progression. *PLoS ONE*, 13(1):e0189887, 2018.

Krumholz, H. M. Big data and new knowledge in medicine: The thinking, training, and tools needed for a learning health system. *Health Affairs*, 33 (7):1163–1170, 2014.

Küffner, R., Zach, N., Norel, R., Hawe, J., Schoenfeld, D., Wang, L., Li, G., Fang, L., Mackey, L., Hardiman, O., et al. Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nature Biotechnology*, 33(1):51, 2015.

Kullback, S. *Information theory and statistics*. 1997.

Laine, S. and Aila, T. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

Lara, O. D. and Labrador, M. A. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials*, 15(3): 1192–1209, 2013.

Lasko, T. A., Denny, J. C., and Levy, M. A. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLoS ONE*, 8(6):e66341, 2013.

Lazo, J. K., Lawson, M., Larsen, P. H., and Waldman, D. M. US economic sensitivity to weather variability. *Bulletin of the American Meteorological Society*, 92(6):709–720, 2011.

Lechner, M. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies*, pp. 43–58. Springer, 2001.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436, 2015.

Lehman, T., Mallireddy, V., Seddon, M., Modali, R., and Ratnasinghe, L. Association between an ABCC8/SUR1 polymorphism and breast cancer risk. *Cancer Research*, 68(9 Supplement):1934–1934, 2008. ISSN 0008-5472. URL `http://cancerres.aacrjournals.org/content/68/9_Supplement/1934`.

Li, C., Xu, K., Zhu, J., and Zhang, B. Triple generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2017.

Libbrecht, M. W. and Noble, W. S. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321, 2015.

Linhardt, L. Learning counterfactual representations for ventilation in critical care: Methods and benchmarks. Master's thesis, ETH Zurich, Switzerland, 2018.

Lipovetsky, S. and Conklin, M. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

Lipscomb, C. E. Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

Lipton, Z. C., Kale, D. C., Elkan, C., and Wetzell, R. Learning to diagnose with LSTM recurrent neural networks. In *International Conference on Learning Representations*, 2016a.

Lipton, Z. C., Kale, D. C., and Wetzel, R. Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series. In *Machine Learning for Healthcare Conference*, pp. 253–270, 2016b.

Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A., and Moroz, I. M. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Biomedical Engineering Online*, 6(1):23, 2007.

Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., and Ramig, L. O. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56(4):1015–1022, 2009.

Liu, Y., Luo, X., Hu, H., Wang, R., Sun, Y., Zeng, R., and Chen, H. Integrative proteomics and tissue microarray profiling indicate the association between overexpressed serum proteins and non-small cell lung cancer. *PLoS ONE*, 7(12):e51748, 2012.

Lo, B. Sharing clinical trial data: Maximizing benefits, minimizing risk. *JAMA*, 313(8):793–794, 2015.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.

Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P. Understanding variable importances in forests of randomized trees. In *Advances in Neural Information Processing Systems*, pp. 431–439, 2013.

Low, S.-K., Kiyotani, K., Mushiroda, T., Daigo, Y., Nakamura, Y., and Zembutsu, H. Association study of genetic polymorphism in ABCC4 with cyclophosphamide-induced adverse drug reactions in breast cancer patients. *Journal of Human Genetics*, 54(10):564, 2009.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4768–4777, 2017.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246, 2017.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518 (7540):529, 2015.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

Mooney, S. J. and Pejaver, V. Big data in public health: Terminology, machine learning, and privacy. *Annual Review of Public Health*, 39:95–112, 2018.

Murdoch, T. B. and Detsky, A. S. The inevitable application of big data to health care. *JAMA*, 309(13):1351–1352, 2013.

Muroi, C., Meier, S., De Luca, V., Mack, D. J., Strässle, C., Schwab, P., Karlen, W., and Keller, E. Automated false alarm reduction in a real-life intensive care setting using motion detection. *Neurocritical Care*, 2019.

Nair, V. and Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In *International Conference on Machine Learning*, pp. 807–814, 2010.

Obermeyer, Z. and Emanuel, E. J. Predicting the future — big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, 375 (13):1216, 2016.

OECD. OECD Regional Statistics (database). *http://stats.oecd.org/ (Accessed 10th Oct 2017)*, 2017.

Oldenhof, H., de Jong, M., Steenhoek, A., and Janknegt, R. Clinical pharmacokinetics of midazolam in intensive care patients, a wide interpatient variability? *Clinical Pharmacology & Therapeutics*, 43(3):263–269, 1988.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. Is object localization for free? Weakly-supervised learning with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 685–694, 2015.

Oresko, J. J., Jin, Z., Cheng, J., Huang, S., Sun, Y., Duschl, H., and Cheng, A. C. A wearable smartphone-based platform for real-time cardiovascular disease detection via electrocardiogram processing. *IEEE Transactions on Information Technology in Biomedicine*, 14(3):734–740, 2010.

Overbeck, T. R., Hupfeld, T., Krause, D., Waldmann-Beushausen, R., Chapuy, B., Güldenzoph, B., Aung, T., Inagaki, N., Schöndube, F. A., Danner, B. C., et al. Intracellular ATP-binding cassette transporter A3 is expressed in lung cancer cells and modulates susceptibility to cisplatin and paclitaxel. *Oncology*, 84(6):362–370, 2013.

Pahwa, R. and Lyons, K. E. Early diagnosis of Parkinson's disease: Recommendations from diagnostic clinical guidelines. *American Journal of Managed Care*, 16(4):94–99, 2010.

Papandreou, G., Chen, L.-C., Murphy, K. P., and Yuille, A. L. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *IEEE International Conference on Computer Vision*, pp. 1742–1750, 2015.

Parikh, R. B., Obermeyer, Z., and Navathe, A. S. Regulation of predictive analytics in medicine. *Science*, 363(6429):810–812, 2019.

Park, S., Shimizu, C., Shimoyama, T., Takeda, M., Ando, M., Kohno, T., Katsumata, N., Kang, Y.-K., Nishio, K., and Fujiwara, Y. Gene expression profiling of ATP-binding cassette (ABC) transporters as a predictor of the pathologic response to neoadjuvant chemotherapy in breast cancer patients. *Breast Cancer Research and Treatment*, 99(1):9–17, 2006.

Partanen, L., Staaf, J., Tanner, M., Tuominen, V. J., Borg, Å., and Isola, J. Amplification and overexpression of the ABCC3 (MRP3) gene in primary breast cancer. *Genes, Chromosomes and Cancer*, 51(9):832–840, 2012.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in PyTorch. *Advances in Neural Information Processing Systems - Autodiff Workshop*, 2017.

Patel, S., Lorincz, K., Hughes, R., Huggins, N., Growdon, J., Standaert, D., Akay, M., Dy, J., Welsh, M., and Bonato, P. Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors. *IEEE Transactions on Information Technology in Biomedicine*, 13(6):864–873, 2009.

Pearl, J. *Causality*. Cambridge University Press, 2009.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: Foundations and learning algorithms*. MIT Press, 2017.

Petersen, R. C., Aisen, P., Beckett, L. A., Donohue, M., Gamst, A., Harvey, D. J., Jack, C., Jagust, W., Shaw, L., Toga, A., et al. Alzheimer's disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology*, 74 (3):201–209, 2010.

Piehler, A. P., Hellum, M., Wenzel, J. J., Kaminski, E., Haug, K. B. F., Kierulf, P., and Kaminski, W. E. The human ABC transporter pseudogene family: Evidence for transcription and gene-pseudogene interference. *BMC Genomics*, 9(1):165, 2008.

Plesinger, F., Klimes, P., Halamek, J., and Jurak, P. Taming of the monitors: Reducing false alarms in intensive care units. *Physiological Measurement*, 37(8):1313, 2016.

Pocevicius, M. Intelligent decision support for diagnosis and monitoring of Parkinson's disease. Master's thesis, ETH Zurich, Switzerland, 2018.

Prasad, N., Cheng, L.-F., Chivers, C., Draugelis, M., and Engelhardt, B. E. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. In *Uncertainty in Artificial Intelligence*, 2017.

Prince, J., Andreotti, F., and Vos, M. D. Multi-Source ensemble learning for the remote prediction of Parkinson's disease in the presence of source-wise missing data. *IEEE Transactions on Biomedical Engineering*, 2018.

Quisel, T., Foschini, L., Signorini, A., and Kale, D. C. Collecting and analyzing millions of mHealth data streams. In *ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, pp. 1971–1980. ACM, 2017.

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C. P., et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Medicine*, 15(11):e1002686, 2018.

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., and Raiko, T. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, pp. 3546–3554, 2015.

Razavian, N., Marcus, J., and Sontag, D. Multi-task prediction of disease onsets from longitudinal laboratory tests. In *Machine Learning for Healthcare Conference*, 2016.

Ribeiro, M. T., Singh, S., and Guestrin, C. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016a.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM, 2016b.

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*, 2018.

Rizzo, G., Copetti, M., Arcuti, S., Martino, D., Fontana, A., and Logroscino, G. Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis. *Neurology*, 86(6):566–576, 2016.

Robins, J. M., Hernan, M. A., and Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

Robson, M. E., Bradbury, A. R., Arun, B., Domchek, S. M., Ford, J. M., Hampel, H. L., Lipkin, S. M., Syngal, S., Wollins, D. S., and Lindor, N. M. American Society of Clinical Oncology policy statement update: Genetic and genomic testing for cancer susceptibility. *Journal of Clinical Oncology*, 33(31):3660–3667, 2015.

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiskỳ, T., and Blunsom, P. Reasoning about entailment with neural attention. In *International Conference on Learning Representations*, 2016.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G. D., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., and Mark, R. G. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Critical Care Medicine*, 39(5):952, 2011.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Saria, S., Rajani, A. K., Gould, J., Koller, D., and Penn, A. A. Integration of early physiological responses predicts later illness severity in preterm infants. *Science Translational Medicine*, 2(48):48ra65–48ra65, 2010.

Schafer, A. The ethics of the randomized clinical trial. *New England Journal of Medicine*, 307(12):719–724, 1982.

Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

Schrag, A., Jahanshahi, M., and Quinn, N. How does Parkinson's disease affect quality of life? A comparison with quality of life in the general population. *Movement Disorders*, 15(6):1112–1118, 2000.

Schuler, A., Baiocchi, M., Tibshirani, R., and Shah, N. A comparison of methods for model selection when estimating individual treatment effects. *arXiv preprint arXiv:1804.05146*, 2018.

Schwab, P. Research plan. ETH Zurich, unpublished manuscript, 2017.

Schwab, P. and Hlavacs, H. Capturing the essence: Towards the automated generation of transparent behavior models. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, pp. 184–190. AAAI, 2015.

Schwab, P. and Karlen, W. CXPlain: Causal explanations for model interpretation under uncertainty. In *Advances in Neural Information Processing Systems (to appear)*, 2019a.

Schwab, P. and Karlen, W. PhoneMD: Learning to diagnose Parkinson's disease from smartphone data. In *AAAI Conference on Artificial Intelligence*, 2019b.

Schwab, P., Scebba, G. C., Zhang, J., Delai, M., and Karlen, W. Beat by beat: Classifying cardiac arrhythmias with recurrent neural networks. In *Computing in Cardiology*, 2017.

Schwab, P., Keller, E., Muroi, C., Mack, D. J., Strässle, C., and Karlen, W. Not to cry wolf: Distantly supervised multitask learning in critical care. In *International Conference on Machine Learning*, 2018a.

Schwab, P., Linhardt, L., and Karlen, W. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*, 2018b.

Schwab, P., Linhardt, L., Bauer, S., Buhmann, J., and Karlen, W. Learning counterfactual representations for estimating individual dose-response curves. *arXiv preprint arXiv:1902.00981*, 2019a.

Schwab, P., Miladinovic, D., and Karlen, W. Granger-causal attentive mixtures of experts: Learning important features with neural networks. In *AAAI Conference on Artificial Intelligence*, 2019b.

Sendelbach, S. and Funk, M. Alarm fatigue: A patient safety concern. *AACN Advanced Critical Care*, 24(4):378–386, 2013.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: Generalization bounds and algorithms. In *International Conference on Machine Learning*, 2017.

Shameer, K., Johnson, K. W., Glicksberg, B. S., Dudley, J. T., and Sengupta, P. P. Machine learning in cardiovascular medicine: Are we there yet? *Heart*, 104(14):1156–1164, 2018.

Shapley, L. S. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, 2017.

Shulman, L. M., Pretzer-Aboff, I., Anderson, K. E., Stevenson, R., Vaughan, C. G., Gruber-Baldini, A. L., Reich, S. G., and Weiner, W. J. Subjective report versus objective measurement of activities of daily living in Parkinson's disease. *Movement Disorders*, 21(6):794–799, 2006.

Silva, R. Observational-interventional priors for dose-response learning. In *Advances in Neural Information Processing Systems*, pp. 1561–1569, 2016.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484, 2016.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations*, 2014.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. SmoothGrad: Removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Smith, M. J., Culhane, A. C., Killeen, S., Kelly, M. A., Wang, J. H., Cotter, T. G., and Redmond, H. P. Mechanisms driving local breast cancer recurrence in a model of breast-conserving surgery. *Annals of Surgical Oncology*, 15(10): 2954–2964, 2008.

Smith, S. L., Gaughan, P., Halliday, D. M., Ju, Q., Aly, N. M., and Playfer, J. R. Diagnosis of Parkinson's disease using evolutionary algorithms. *Genetic Programming and Evolvable Machines*, 8(4):433–447, 2007.

Soucek, P., Hlavac, V., Elsnerova, K., Vaclavikova, R., Kozevnikovova, R., and Raus, K. Whole exome sequencing analysis of ABCC8 and ABCD2 genes associating with clinical course of breast carcinoma. *Physiological Research*, 64:S549, 2015.

Springenberg, J. T. Unsupervised and semi-supervised learning with categorical generative adversarial networks. In *International Conference on Learning Representations*, 2016.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Srivastava, R. K., Greff, K., and Schmidhuber, J. Training very deep networks. In *Advances in Neural Information Processing Systems*, pp. 2377–2385, 2015.

Stollenga, M. F., Masci, J., Gomez, F., and Schmidhuber, J. Deep networks with internal selective attention through feedback connections. In *Advances in Neural Information Processing Systems*, pp. 3545–3553, 2014.

Stone, R. The assumptions on which causal inferences rest. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 455–466, 1993.

Štrumbelj, E., Kononenko, I., and Šikonja, M. R. Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering*, 68(10):886–904, 2009.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3):e1001779, 2015.

Suetens, P. *Fundamentals of medical imaging*. Cambridge University Press, 2017.

Suhara, Y., Xu, Y., and Pentland, A. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *International Conference on World Wide Web*, 2017.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017.

Szakács, G., Annereau, J.-P., Lababidi, S., Shankavaram, U., Arciello, A., Bussey, K. J., Reinhold, W., Guo, Y., Kruh, G. D., Reimers, M., et al. Predicting drug sensitivity and resistance: Profiling ABC transporter genes in cancer cells. *Cancer Cell*, 6(2):129–137, 2004.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence*, 2017.

Teh, Y., Bapst, V., Pascanu, R., Heess, N., Quan, J., Kirkpatrick, J., Czarnecki, W. M., and Hadsell, R. Distral: Robust multitask reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 4497–4507, 2017.

Tsanas, A., Little, M. A., McSharry, P. E., and Ramig, L. O. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57(4):884–893, 2010.

Tsanas, A., Little, M. A., McSharry, P. E., and Ramig, L. O. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful

quantification of average Parkinson's disease symptom severity. *Journal of the Royal Society Interface*, 2011.

Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., and Ramig, L. O. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 59(5): 1264–1271, 2012.

Van Ree, J. H., Jeganathan, K. B., Malureanu, L., and Van Deursen, J. M. Overexpression of the E2 ubiquitin–conjugating enzyme UbcH10 causes chromosome missegregation and tumor formation. *Journal of Cell Biology*, 188(1):83–100, 2010.

Vayena, E., Blasimme, A., and Cohen, I. G. Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, 15(11):e1002689, 2018.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, pp. 1096–1103, 2008.

Vos, T., Allen, C., Arora, M., Barber, R. M., Bhutta, Z. A., Brown, A., Carter, A., Casey, D. C., Charlson, F. J., Chen, A. Z., et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: A systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, 388(10053):1545–1602, 2016.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 2017.

Wang, L., Alpert, K. I., Calhoun, V. D., Cobia, D. J., Keator, D. B., King, M. D., Kogan, A., Landis, D., Tallis, M., Turner, M. D., et al. SchizConnect: Mediating neuroimaging databases on schizophrenia and related disorders for large-scale integration. *NeuroImage*, 124:1155–1167, 2016.

Wang, Y., Hao, D.-P., Li, J.-J., Wang, L., and Di, L.-J. Genome-wide

methylome and chromatin interactome identify abnormal enhancer to be risk factor of breast cancer. *Oncotarget*, 8(27):44705, 2017.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

Wiens, J., Guttag, J., and Horvitz, E. Patient risk stratification with time-varying parameters: A multitask learning approach. *Journal of Machine Learning Research*, 17(1):2797–2819, 2016.

Xu, J., Schwing, A. G., and Urtasun, R. Learning to segment under various forms of weak supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3781–3790, 2015a.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pp. 2048–2057, 2015b.

Yaeger, K. A., Martini, M., Yaniv, G., Oermann, E. K., and Costa, A. B. United States regulatory approval of medical devices and software applications enhanced by artificial intelligence. *Health Policy and Technology*, 8(2):192–197, 2019.

Yamada, A., Ishikawa, T., Ota, I., Kimura, M., Shimizu, D., Tanabe, M., Chishima, T., Sasaki, T., Ichikawa, Y., Morita, S., et al. High expression of ATP-binding cassette transporter ABCC11 in breast tumors is associated with aggressive subtypes and low disease-free survival. *Breast Cancer Research and Treatment*, 137(3):773–782, 2013.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. J., and Hovy, E. H. Hierarchical attention networks for document classification. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489, 2016.

Yeatman, T., Centeno, B. A., and Bloom, G. C. Hybrid model for the classification of carcinoma subtypes, June 16 2015. US Patent 9,057,108.

Yoon, J., Jordon, J., and van der Schaar, M. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

Zaske, D. E., Irvine, P., Strand, L. M., Strate, R. G., Cipolle, R. J., and Rotschafer, J. Wide interpatient variations in gentamicin dose requirements for geriatric patients. *JAMA*, 248(23):3122–3126, 1982.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833, 2014.

Zeng, D., Liu, K., Chen, Y., and Zhao, J. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Conference on Empirical Methods in Natural Language Processing*, pp. 1753–1762, 2015.

Zhan, A., Little, M. A., Harris, D. A., Abiola, S. O., Dorsey, E., Saria, S., and Terzis, A. High frequency remote monitoring of Parkinson's disease via smartphone: Platform overview and medication response detection. *arXiv preprint arXiv:1601.00960*, 2016.

Zhan, A., Mohan, S., Tarolli, C., Schneider, R. B., Adams, J. L., Sharma, S., Elson, M. J., Spear, K. L., Glidden, A. M., Little, M. A., et al. Using smartphones and machine learning to quantify Parkinson disease severity: The mobile Parkinson disease score. *JAMA Neurology*, 2018.

Zhang, H., Cao, J., Li, L., Liu, Y., Zhao, H., Li, N., Li, B., Zhang, A., Huang, H., Chen, S., et al. Identification of urine protein biomarkers with the potential for early detection of lung cancer. *Scientific Reports*, 5:11805, 2015.

Zhang, Z., Xie, Y., Xing, F., McGough, M., and Yang, L. MDNet: A semantically and visually interpretable medical image diagnosis network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Zheng, B., Yoon, S. W., and Lam, S. S. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. *Expert Systems with Applications*, 41(4):1476–1482, 2014.