Deep learning for health outcome prediction

Arinbjörn Kolbeinsson

Thesis submitted for the degree of
Doctor of Philosophy

December 2020

Department of Epidemiology and Biostatistics
Imperial College London

Abstract

Modern medical data contains rich information that allows us to make new types of inferences to predict health outcomes. However, the complexity of modern medical data has rendered many classical analysis approaches insufficient. Machine learning with deep neural networks enables computational models to process raw data and learn useful representations with multiple levels of abstraction. In this thesis, I present novel deep learning methods for health outcome prediction from brain MRI and genomic data. I show that a deep neural network can learn a biomarker from structural brain MRI and that this biomarker provides a useful measure for investigating brain and systemic health, can augment neuroradiological research and potentially serve as a decision-support tool in clinical environments. I also develop two tensor methods for deep neural networks: the first, tensor dropout, for improving the robustness of deep neural networks, and the second, Kronecker machines, for combining multiple sources of data to improve prediction accuracy. Finally, I present a novel deep learning method for predicting polygenic risk scores from genome sequences by leveraging both local and global interactions between genetic variants. These contributions demonstrate the benefits of using deep learning for health outcome prediction in both research and clinical settings.

Licence

Statement of Originality

The work presented in this thesis is the result of my own research. Any work which is the outcome of collaboration is acknowledged and referenced in the text.

Acknowledgements

I thank my incredible team of supervisors, **Joanna Tzoulaki**, **Sarah Filippi**, **Yannis Panagakis** and **Paul Elliott**.

I am very grateful to **Abbas Dehghan** and **Paul Matthews**, who have both been instrumental over the course of my PhD and accelerated my development as a researcher. **Korbinian Strimmer**, for his supervision during the first year of my PhD and for having the faith in me to pursue research in deep learning in health.

Jean Kossaifi, for our collaborations. Without his support and advice, I would not be the machine learning researcher I am today. **Maja Pantic** and **Andrew Blake**, for the opportunity to intern at Samsung AI in Cambridge where I worked with an amazing team of researchers.

Hideaki Suzuki, for sharing his valuable knowledge on brain modelling, **Wenjia Bai**, for his expertise on UK Biobank MRI data and **He Gao**, for her guidance on GWAS and genotype data.

Deborah Schneider-Luftman for the numerous engaging discussions on UK Biobank data, and her advice on how to efficiently extract it. **Pyry Helkkula** for his advice and deep insight on genomic data. **Kai Arulkumaran** and **Pierre Richemond**, for organising the Deep Learning Reading Group at Imperial which broadened my horizon and brought me ideas and inspiration every week.

I thank my examiners, **Seth Flaxman** and **John Gallacher**, for lending their expertise to examine my work. It was a thought-provoking experience and I thoroughly enjoyed our insightful discussions.

I thank the **Medical Research Council** for funding this work.

Last, but not least, I thank my parents and my brother for their support and encouragement. Without them, this research would never have been realised.

Contents

Acronyms

BMI Body Mass Index.

CAD Coronary Artery Disease.

CNN Convolutional Neural Networks.

CP Canonical Polyadic.

CT Computed Tomography.

FIS Fluid Intelligence Score.

fMRI Functional Magnetic Resonance Imaging.

GPU Graphics Processing Unit.

GWAS Genome Wide Association Study.

HES Hospital Episode Statistics.

IDP Image Derived Phenotype.

LD Linkage Disequilibrium.

LSTM Long Short-Term Memory.

MAE Mean Absolute Error.

MLP Multi-Layer Perceptron.

MR Magnetic Resonance.

MRI Magnetic Resonance Imaging.

MSE Mean Squared Error.

PGS Polygenic Score.

R-TRL Randomised Tensor Regression Layer.

ReLU Rectified Linear Unit.

Resnet Residual Network.

RNN Recurrent Neural Networks.

SBP Systolic Blood Pressure.

SGD Stochastic Gradient Descent.

SNP Single Nucleotide Polymorphisms.

SVD Singular Value Decomposition.

TRL Tensor Regression Layer.

UMAP Uniform Manifold Approximation and Projection.

List of Figures

List of Tables

Chapter 1

Introduction

As more fundamental and unprocessed biomedical data is collected, the task of feature selection and representation generation shifts away from the human analyst towards machine learning systems. These modern unprocessed data are characterised by rich intrinsic structure, whose properties derive from high-order nonlinear underlying biological mechanisms from which the data are captured. For example, information in medical images is defined not only by the intensities of individual pixels, but also by their spatial arrangement which gives rise to patterns of interest. Learning to recognise minute differences in patterns between thousands of images and genome sequences, each made up of millions of features, is not feasible without computational methods.

Unfortunately, this structural complexity of raw biomedical data has rendered many classical analysis approaches obsolete. Univariate methods often struggle to find relationships in large, structure-rich, multi-source data sets (Angermueller et al. 2016). Moreover, many methods used in practice today were designed when data were scarce, practitioners collected few features and observations in small case-control studies. Today, the average individual in a developed country will generate approximately 1.2 TB of health data in their lifetime (Nature Editorial 2016). These data are high dimensional, rich in structure, multi-modal, multi-label and inevitably noisy. There is a clear need for methods that are able to make intelligent inferences from large medical data sets. Using knowledge on structural priors, we can build models that learn better representations using richer signals from the data.

Other fields with abundant amounts of data, such as e-commerce, online advertising and robotics, have been revolutionised in the last decade by extensive use

of machine learning, in particular deep learning. However, as we shall see, simply porting methods over to healthcare is inadequate. For a machine learning system to be used in medical applications, all its aspects, including experimental setup, models and training regime, need to be adapted to the data and task which is being learnt. Systems used in healthcare serve critical applications and therefore need to be accurate, robust to samples that are outside of the training distribution and interpretable if used in certain clinical settings. The applications of deep learning in medicine are only beginning to be realised, but their value is clear.

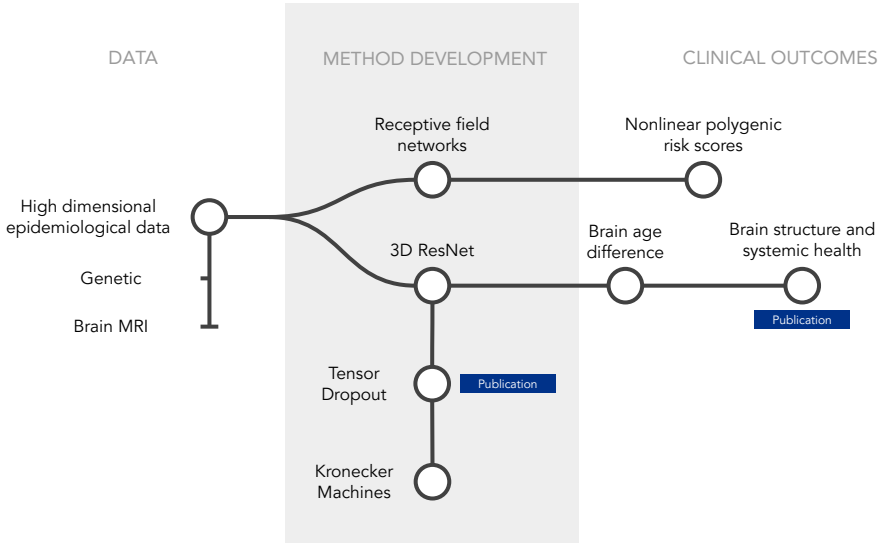


Figure 1.1: An overview of the work presented in this thesis. The left column represents the primary data used as input for learning: brain MRI and genome sequences. The centre column describes the methodological contributions. Finally, the right column contains the health outcomes that can be derived by applying these methods to the respective data.

In this thesis I present my work on deep learning for health outcome prediction, specifically on Magnetic Resonance (MR) imaging and genomic data. Figure 1.1 gives an overview of the work, linking together the data sources, methodological contributions and clinically-relevant health outcomes. The ultimate aim of these predictions is to stratify individuals who would benefit from treatment or risk remediation. This requires relevant data about the individual, a way of interpreting the information and transforming it into relevant actions.

I build on recent discoveries and studies to:

- Develop a 3D-Residual neural network that achieves state-of-the-art results on predicting clinically relevant traits from MR imaging (Chapter 3).
- Demonstrate that 3D-Residual networks can learn a biomarker from brain

structural MRI, and show that this biomarker is associated with with cardiometabolic and cognitive diseases. I explore the causality of the associations and assess the relative contributions of features from different brain regions to identify those that were most informative for the model (Chapter 4).

- Develop a tensor learning method: tensor dropout, which gives improved robustness and generalisability to deep neural networks. Tensor dropout stochastically drops ranks of decomposed tensor weights. I demonstrate the application of this approach on medical imaging tasks and perform extensive hyperparameter analysis (Chapter 5, section 2).

- Introduce a method for combining multi-modal data for health outcome prediction. The method learns interactions between separate inputs using a factorised Kronecker product to reduce the number of parameters. This constrains the learning space and leads to improved learning properties. I demonstrate this by learning tasks from T1 and T2-weighted brain MRI jointly (Chapter 5, section 3).

- Present Receptive field networks, a novel deep learning architecture for polygenic risk score prediction 6. The approach leverages local structure in the genome and models nonlinear interactions between variants. In a study using UK Biobank data to predict coronary artery disease from genotype data, receptive field networks performed comparably to state-of-the-art linear methods. Neural layer activation analysis demonstrates proof-of-concept interpretability of model predictions.

Chapter 2

Deep learning for health

In this chapter I review relevant literature and methodology to provide background for following chapters. In his seminal paper (Rosenblatt 1958) on information storage and organisation of the brain, Rosenblatt stated three fundamental questions to be answered in order to understand learning and thinking of complex organisms. Although the purpose of this is work is not to accurately model biological cognition, the questions set out a general framework of artificial intelligence and learning:

1. How is information learned[1]?
2. How is this information stored, remembered or modelled?
3. How does information in memory dictate actions?

The three are intricately connected; it is obvious that the manner in which information can be learned is dictated by how it is stored. In this thesis, I focus explicitly on the second of these questions and implicitly on the first and third. In mathematical notation, this can be elegantly described as:

$$\mathbf{y} = \boldsymbol{\phi}(\boldsymbol{\theta}, \mathbf{x})$$

where $\boldsymbol{\phi}$ is the *model* that defines the operations and processes that operate on the input x. $\boldsymbol{\theta}$ are the parameters of the model and are learned by an optimisation algorithm. The output $\mathbf{y}$ can take many forms depending on the task. Taking this

[1]Rosenblatt's first question was "How is information about the physical world sensed, or detected, by the biological system?" which he claimed was largely solved by sensory physiology. Here, I adapt this idea to the context of artificial intelligence.

all together, a popular definition of *machine learning* was set forth by Mitchell et al. (1997) and states that "a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

I will begin by describing one way to model intelligence: deep neural networks (section 2.1). Then, I overview algorithms used for training neural networks, or in essence: how to learn (section 2.2). Finally, as this thesis is concerned with the application of machine learning in a healthcare and biomedicine, I review the domain-specific literature (section 2.3).

2.1 Deep neural networks

The concept of (artificial) neural networks traces its roots to the 1950s, but their application did not reach mainstream usage until the advent of a) a suitable training algorithm (backpropagation) b) efficient computation (e.g. GPUs) and c) large datasets (on the scale of e.g. UK Biobank). In this section I will describe the building blocks of deep neural networks and overview the most common architectures in use today.

2.1.1 The perceptron

The fundamental processing unit of all neural networks is the perceptron. Originally proposed by Rosenblatt (1958), building on earlier learning ideas by Hebb (1949), these units were inspired by way in which higher organisms store and use observed information. Combined with a nonlinear activation function $\mathcal{F}$, the processing unit becomes highly versatile and greater than the sum of its parts (Cybenko 1989). The perceptron operation is as follows:

$$\boldsymbol{y} = \mathcal{F}\Big(\sum_i \boldsymbol{w_i x_i} + b\Big)$$

here, x_i is the perceptron's input and w_i are the perceptron's weights for input i, and b an optional bias parameter. The choice of nonlinear activation function $\mathcal{F}$ is dependent on the model structure and training objective. Popular choices include tanh, the sigmoid function and, in particular, the Rectified Linear Unit (ReLU) as its derivative is easily computed.

2.1.2 Modelling: Deep neural networks

Individual perceptrons are insufficient for modelling complex systems. Connecting many such units together in a directed (feed-forward) network creates a Multi-Layer Perceptron (MLP), a type of *neural network*, as illustrated in figure 2.1. The first layer of these networks is shown here to represent the inputs (data). Hidden layers encode abstract representations of the input that are optimised to learn a desired output. The output, and subsequently the output layer structure, is dependent on the task: single unit for regression tasks, C units for classification tasks with C classes, or a custom shape that is appropriate for the given task.

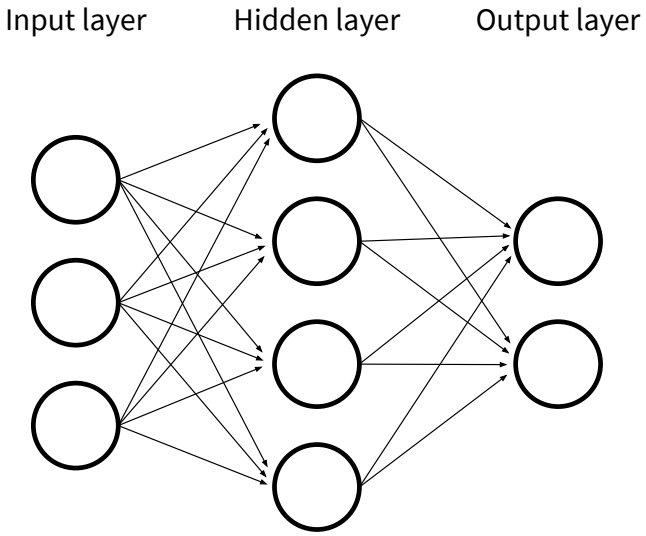


Figure 2.1: Multi-layer perceptron. Each unit represents a perceptron, with arrows as the inputs that are multiplied by learned individual weight parameters. The bias terms have been omitted for simplicity.

The number of hidden layers, or the depth, of a neural network is variable. In computer vision deep neural networks with over 1 000 layers are not unheard of (He et al. 2016), while networks for sequential data can have as few as three layers, yet achieve state-of-the-art results[2].

Neural networks are powerful function approximators but require relatively large datasets for accurate representations to be learnt. To combat this, domain knowledge is used to design structural inductive biases that constrain the architecture and limit the expressiveness of the network. Examples include the Convolutional Neural Networks (CNN) for images, the Recurrent Neural Networks (RNN) for sequential data and graph neural networks for network data.

[2] A single-layer recurrent neural network operating on a sequence of length T practically become networks of depth T

2.1.3 Convolutional neural networks

Images usually[3] encode only relative relationships between objects, e.g. traffic lights can be characterised by the relative red-yellow-green colour arrangement, but the location, orientation and scale of the lights within the image can be arbitrary. Furthermore, objects or other motifs, patterns or regions of interest are often repeated within the same image, motivating representations that can be learned jointly and shared across different parts of the image. Space-invariant neural networks, such as the neocognitron (Fukushima and Miyake 1982) are the result of this incentive. Further model developments and improvements to training (back-propagation) led to the modern day CNN implementation (LeCun, Boser, et al. 1989).

A CNN is a network with convolutional layers. The convolutional operator involves sliding a kernel matrix across the input image and multiplying elementwise at each indexed location. The size of the kernel matrix and the locations are controlled by the kernel size and stride size parameters, respectively.

The operation is defined as:

$$S_{i,j,k} = \sum_{l,m,n} I_{i+m-1,j+n-1,l} \cdot K_{k,m,n,l}$$

where S is the output of I, an image with two spatial dimensions, indexed by (i, j), and channels l, convolved with a filter K, indexed by (k, m, n, l). k is the index of the output channels. To complete the layer, a bias is added to the output and passed through a nonlinear activation function. Stacking multiple such layers together creates a CNN.

Technically, this is the *cross-correlation* operation, as implemented by most deep learning libraries. However, since the indexing of weight kernels is reversible, i.e. it does not matter whether they are indexed over the image left/top to right/bottom or in reverse the two operations become practically equivalent as long as it is consistent throughout training and testing (I. Goodfellow et al. 2016).

With this convolutional operation, the size of the image remains constant through the network. Although distant regions in the input image will eventually be convolved together as deeper layers connect adjacent representations from previous layers. Connecting regions in this manner is both slow and inefficient. Pooling

[3]In highly controlled setups, such as medical imaging, the absolute location of objects can be managed so that their location and orientation is approximately standardised. However, this does not usually translate to pixel-level precision of object placement.

layers have been a popular mitigation option. Pooling layers downsample the image, at any depth of the CNN, to create a smaller and deterministic representation. Max-pooling and average-pooling is defined as the maximum or average, respectively, pixel values are selected from a rolling filter, typically of size 2×2. This summarises the image but loses the information from the unselected outputs. State-of-the-art architectures make very limited uses of pooling layers and instead perform downsampling by increasing the stride of the kernel.

2.1.4 Other neural network architectures

While all neural networks are built using the same principles and building blocks, its structural properties can be configured to promote certain structural priors (inductive biases), like we have seen in CNNs for image and grid-like data. Here, I will briefly describe a selection of popular neural networks architectures and the data they are designed to work with. For a comprehensive review of deep neural networks please refer to the work of LeCun, Bengio, and G. Hinton (2015).

Recurrent neural networks. Data ordered in a sequence, in particular a one-directional sequence such as temporal data, is easily processed using recurrent neural networks. At timestep t these networks use the output of the previous timestep, $t-1$, as an additional input. $\mathbf{h}_t = \mathcal{F}(\mathbf{h}_{t-1}, \mathbf{x}_t)$ This makes them applicable to sequences of arbitrary length. The concept of Long Short-Term Memory (Hochreiter and Schmidhuber 1997) greatly improves recurrent neural networks by avoiding vanishing gradients which causes information in the early part of a sequence to be "forgotten". These networks are highly suitable for learning mappings from one sequence to another (Sutskever, Vinyals, and Le 2014), for example in language translation.

Graph neural networks. Not all data has Euclidean structure. Many data lie on manifolds or are represented as a set of nodes connected in a graph. For such data, rather than mapping the manifold onto a Euclidean space and applying CNNs, the aptly-named class of graph neural networks can learn directly from the data and leverage the inherent structural properties (Bronstein et al. 2017). These networks have been used for visual question answering (Norcliffe-Brown, Vafeias, and Parisot 2018), where answers to written questions have to be inferred from structural relationships in an image.

2.2 Learning: Backpropagation, loss and optimisation

In the previous section, we have seen what to learn; but only modelling the structure of intelligence and defining the operations is not sufficient. In this section we will explore *how to learn*. Our definition of machine learning from Chapter 1 is, "A model is said to learn from data if its performance on a given task improves after the data is taken into account" (Mitchell et al. 1997). The task needs to be defined to determine whether performance is improving. In supervised learning, the performance on the task is defined *a priori* as a loss function $\mathcal{L}$ (also known as an objective function). The loss can then be backpropagated through the network to calculate the parameter updates required to minimise the loss. These parameter updates are given as gradient estimations from an optimisation algorithm. In this section, I will overview these three properties: loss functions, backpropagation and optimisation, and describe how they work together to train neural networks.

2.2.1 Loss functions

A loss function quantifies the fit of the model to the data. By convention, the objective is to minimise the loss function so that $\mathcal{L} = 0$. The choice of the loss function depends on the problem and should reflect the task.

For regression tasks, this is often the mean squared error (MSE). In the supervised setting with training pairs (X, Y), it is defined as

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

where Y_i is the true label of the ith sample, $\hat{Y}_i = \phi(\theta, X_i)$ the label value as predicted by the model and N is the number of samples.

For classification tasks, loss is often measured as the cross entropy between the true and predicted distributions:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N} Y_i \log(\hat{Y}_i)$$

2.2.2 Backpropagation and optimisation

Given a model with a set of parameters θ and a loss function $\mathcal{L}$ to describe the fit to the model, we now require a way to update the model parameters to better fit the data. This is accomplished with Stochastic Gradient Descent: repeatedly fitting mini-batches, B (a small set of samples from the training data, also known as simply 'batch'), and computing the gradients of every parameter by backpropagation (Linnainmaa 1970; Rumelhart, Geoffrey E Hinton, and Williams 1986). Parameter updates are then given by $\theta = \theta - \nu \nabla_\theta \mathcal{L}_B$.

Extensions to SGD include Adam (Kingma and Ba 2015), a popular optimiser for deep models, that uses adaptive estimates of moments (mean and uncentred variance of the gradients) by taking previous gradients into account (with a moving average). Adagrad (McMahan and Streeter 2010; Duchi, Hazan, and Singer 2011) and RMSprop (Tieleman and G. Hinton 2012) are common alternatives.

2.2.3 Training hyperparameters

Learning rate defines the size of individual parameter updates in the gradient direction that minimises the loss. Larger learning rates reduce the number of updates needed to reach areas of lower loss, but will perpetually "overshoot" shallow minima. It is common practice to reduce the learning rate as training progresses through scheduled learning rate reductions at pre-defined epochs. Optimisers with adaptive momentum (e.g. Adam) also help with this.

Regularisation is a class of tools and practices used for improving the generalisability of models. Neural networks are often over-parameterised and have capacity to overfit to the training data. To alleviate this, *weight decay* (also known as shrinkage) can be applied to the parameters of neural networks (or any parameterised model). This penalises over reliance on small numbers of parameters and encourages the model to learn more robust representations. Since the weight penalisation is a component in the overall loss being minimised, its level can be controlled by modifying the proportion of the loss corresponding to weight decay. This is done by multiplying the weight decay term with a scalar value. *Dropout* (Srivastava et al. 2014), used primarily for neural networks, operates in a similar way, but instead of modifying the loss term, the model architecture is changed so that units have a probability p of being zero during training. This forces the network to utilise its capacity and not become over reliant on a small number of units. p is a tunable hyperparameter. *Early stopping* is a heuristic trick that works by ending the train-

ing process before the network has the opportunity to learn details that are specific to the training set samples. Finally, batch normalisation, or *Batchnorm*, (Ioffe and Szegedy 2015) can be applied between layers to normalise the signal, per mini-batch. Note batch-norm is applied after any layer-specific activation functions.

2.3 Medical applications of deep learning

The qualities of deep learning lend themselves to many medical applications. Deep learning methods excel in pattern recognition and are notably robust to different types of noise or artefacts that are irrelevant to the defined objective. Deep neural networks have achieved state-of-the-art performance on diagnosis tasks, particularly in medical imaging.

Early disease identification and intervention can lead to better treatment and improved prognosis. Recent studies have demonstrated the application of CNNs to recognition of Alzheimer's disease from MRI scans (Sarraf and Tofighi 2016; Islam and Yanqing Zhang 2017). Deep learning has been applied to histopathologic cancer diagnosis (Litjens et al. 2016; Bayramoglu, Kannala, and Heikkilä 2016) and classification of skin cancer based on images Esteva, Kuprel, et al. 2017. The latter having implications for large-scale deployment on mobile devices, if the model is robust to different lighting environments. Medical image segmentation is an important processing step in diagnosis, radiotherapy and research. The task consists of assigning pixels (or voxels in 3D imaging) to connected regions that have a clinical association. UNet (Ronneberger, Fischer, and Brox 2015) revolutionised the field with clever use of skip connections between learned upsampled representations and downsampled versions of the image using a matching downsampling layer. The UNet (and its derivatives) have been used extensively with excellent performance on tumour segmentation (X. Li et al. 2018).

Impact in medicine outside of imaging has been less predictable. Some fields suffer from a lack of either labelled data or a suitable architecture to match the data structure, i.e. what CNNs do for images and RNNs for sequence data. Recent implementations have demonstrated the use of RNNs for classification of myocardial infarction from (sequential) ECG (Acharya et al. 2017; Baloglu et al. 2019). Other deep learning models have predicted in-hospital mortality, unplanned readmission, prolonged length of stay and final discharge diagnoses with greater accuracy than clinically-used methods (Rajkomar et al. 2018). In genetics, applications include regulatory genomics (Zhuang, Shen, and Pan 2019) and identification of variants

from short-read sequence data (Poplin, Chang, et al. 2018).

In addition to serving as a direct replacement for existing approaches, deep learning promises to enabling completely new types of analyses, the impacts of which are only beginning to be hypothesised. Protein folding is considered by many as one of the fundamental challenges in science and far from being solved. This year, AlphaFold (Senior et al. 2020) achieved unprecedented prediction scores for protein folding, leading to hopes that the challenge might be, at least partly, solved. The practical implications remain uncertain, but deep neural networks have clearly established themselves as a versatile modelling technique that demonstrate state-of-the-art results on a wide range of tasks.

Certain limitations of deep learning approaches inhibit their deployment in healthcare, or limit it completely to more abstract medical research. The outputs of deep learning models are notoriously hard to interpret and explain, giving them the nickname of "black-box" models. The theory of interpreting deep learning, specifically deep neural networks, is still in its infancy but tools have been introduced to explain the predictions or decision that a model makes. These tools are not limited to the domain of medical applications but can be restricted to a certain class of models, e.g. Quantitative Testing with Concept Activation Vectors (TCAV) (Kim et al. 2018) is applicable to neural networks while Shapley values (Lundberg and S.-I. Lee 2017) work with any parameterised prediction model.

The entire field of deep learning in medicine is too vast to be covered in this thesis. Instead, relevant developments and models are discussed in subsequent chapters. I refer the reader to comprehensive reviews by Angermueller et al. (2016), Min, B. Lee, and Yoon (2017) and Esteva, Robicquet, et al. (2019).

Chapter 3

3D-Residual networks for medical imaging

In this chapter I present my work towards building better systems that can learn rich representations from medical data, specifically brain structural MRI. The motivation is to improve decision-making in healthcare. For this to happen, relevant information must be analysed, interpreted and acted on. The 3D-Resnet model introduced here aims to perform the first two of those tasks.

3.1 Introduction

Medical imaging allows for detailed and non-invasive visual capture of internal body structure and mechanisms. MRI, Computed Tomography (CT) (radiography) and ultrasound are used at most major hospitals and medical research centres around the world. The data generated by these techniques is high-dimensional (on the order of millions). Relationships between dimensions, such as local and global spatial structure, and time sequence in fMRI, also carry important information. These structural relationships are intricate and difficult to characterise due to low signal-to-noise ratios. Indeed, regressing feature-engineered Image Derived Phenotype (IDP), or other classical methods, often fail at capturing nonlinear relationships (Akkus et al. 2017). Recently, CNNs have demonstrated state-of-the-art performance on many medical imaging tasks, including brain segmentation, estimating bone fractures (C.-T. Cheng et al. 2019), identifying lung nodules (Massion et al. 2020) and diabetic retinopathy (Gulshan et al. 2016).

However, porting computer-vision models over from natural images to medical imag-

ing is not trivial. MRI data has particular attributes that can be considered to maximise information gain and predictive performance. Most methods do not leverage 3D structure in the data and use only 2D slices, thereby breaking important structural relationships. Furthermore, finding the appropriate CNNs optimisation strategies for MRI-based learning is a major challenge, leading to suboptimal prediction accuracy of many models. This is partly due to computational complexity and lack of computational resources.

In this chapter I will present 3D-Residual networks: an architecture that extends general computer-vision methods to achieve state-of-the-art performance for learning from MRI data. This complements similar 3D CNNs (Korolev et al. 2017), which were developed concurrently. Specifically, the contributions of this chapter are:

- The development of a CNN with 3D spatial convolutions that leverages the full structure of the input volumes. The network makes use of residual connections to allow deeper and richer representations to be learned more effectively.
- To demonstrate the performance of the model in the context of baseline methods on four different brain MRI tasks.
- An ablation study to explore the properties of the model and its sensitivity to hyperparameters.

3.2 Data: UK Biobank

UK Biobank is a population-based cohort study of $\sim$ 500 000 participants, who were recruited from the UK general population between 2006 and 2010. At baseline, participants, who were between 40-69 years old, provided blood samples for biochemical tests and genotyping and a wide range of self-reported information and physical measurements, and consented for their data to be linked to Hospital Episode Statistics (HES). An overview of selected information available in UK Biobank is shown in table 6.2. Detailed protocols for obtaining the measurements from participants have been described in the literature (Sudlow et al. 2015a). An imaging extension to the existing UK Biobank study was initiated in 2016 with plans to scan 100 000 individuals from the cohort by 2022-23 (K. L. Miller et al. 2016).

The initial release available to researchers at the start of this project contained images for 5 000 individuals and was increased in stages to the currently available set

of 21 382 individuals. MRI modalities in this release include: T1-weighted structural images, T2-weighted (FLAIR) structural images, diffusion imaging, resting-state and task functional imaging, and diffusion imaging. In this thesis, T1 and T2-weighted images were used. T1-weighted images emphasise contrast between grey and white matter. T2-weighted (FLAIR) images emphasise water-rich tissue, this is of particular interest when investigating pathology, such as lesions (S. Smith, Almagro, and K. Miller 2017). All images were captured at UK Biobank imaging centres on identical 3 T Siemens Skyra scanners (software platform VD13). Each image (both T1 and T2-weighted) was single channel with dimensions $182 \times 218 \times 182$ at $1\,\mathrm{mm}^3$ resolution. Available reconstructed images were aligned to the MNI152 template (Jenkinson et al. 2002).

Although all data were collected at baseline, many measurements, including weight, blood pressure and cognitive function scores, are being repeated for longitudinal analysis. Importantly, follow-up images are currently being captured for participants in the imaging sub-cohort approximately 5 years after initial image capture. This enables investigations into temporal changes as captured by imaging and prospective studies by looking at downstream outcomes.

In this thesis, I am primarily interested in understanding the UK Biobank structural brain imaging and genomic data, and how it can be used to infer and predict health on an individual level.

3.3 Three-dimensional Residual Networks

Here I describe the three-dimensional (3D) CNN that predicts phenotypes from brain structural MRI data. I build on Resnets (He et al. 2016) and leverage its residual connections to construct deeper three-dimensional networks.

The network architecture is divided into three processing stages, which are illustrated in figure 3.1. The first stage consists of a 3D-convolutional layer, batch-normalisation (Ioffe and Szegedy 2015), ReLU activation and max-pooling. In this implementation, the 3D-convolutions have kernel size $7 \times 7 \times 7$ and a stride of 1. The input to the first convolutional layer is a single-channel image and its output a 64-channel activation tensor with three spatial dimensions. The max-pooling operation downsamples the tensor with a window size of 3 and a stride of 2.

The second stage contains four 3D-residual blocks that learn abstract representations of the image. Each residual block has two sets of 3D convolutions (with kernel

Table 3.1: Overview of UK Biobank data availability in 2020. Note that this is not a complete list and that not all of the data listed here is used in this thesis.

Feature	Sample size	Dimensions	Description
T1-weighted brain MRI	19 379	$182 \times 218 \times 182$	3D image with 1mm^3 resolution.
T2-weighted brain MRI	19 379	$182 \times 218 \times 182$	3D image with 1mm^3 resolution.
Resting-state brain fMRI	19 379	$88 \times 88 \times 64 \times 490$	fMRI captured for 6 minutes with 490 time points. Individual in resting cognitive state.
Active-task brain fMRI	19 379	$88 \times 88 \times 64 \times 332$	fMRI captured over 332 time points with individual performing a cognitive task.
Genome sequences	500 000	Up to $3 \times 93\,095\,623$	Multiple versions are available including genotyped and imputed data, with or without coding variants etc.
ICD-10 diagnoses	500 000	$\mathbf{x} \in \mathbb{B}^{1\times 19\,154}$	Diagnosis from electronic health records.
Biological information (continuous)	500 000	$\mathbf{x} \in \mathbb{R}^{1\times 950}$	Data including age, weight, blood pressure, cognitive test scores. Available at three time points.
Biological information (binary classes)	500 000	$\mathbf{x} \in \mathbb{B}^{1\times 138}$	Data including sex, attempted cognitive test. Available at three time points.
Biological information (categorical)	500 000	$\mathbf{x} \in \mathbb{N}^{1\times 31}$	Data including education level, smoking status, exercise per week. Available at three time points.

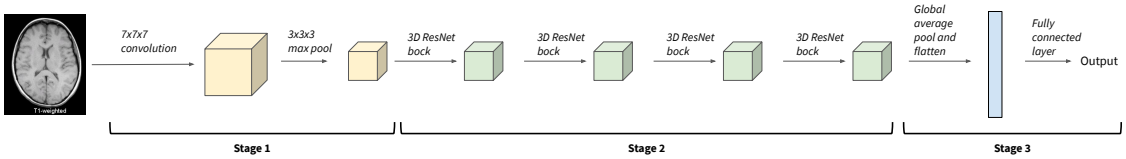


Figure 3.1: Schematic overview of 3D-ResNet for MRI analysis. The architecture can be divided into three processing stages for illustrative purposes. It is trained end-to-end. The network extends the residual block architecture to three spatial dimensions and maintains full spatial structure throughout the convolutional layers.

size $3\times3\times3$), batch normalisation and ReLU activation. Residual connections are a type of skip connection that help the parameter gradient updates to back-propagate through the network. Gradient information can bypass non-linear layers to reduce the effect of vanishing gradients. Downsampling through the residual blocks is controlled entirely by striding the convolutions by two voxels, instead of a pooling operation. Each of the four blocks generates a more abstract representation of the structural image that is more compact in spatial dimensions with more capacity in the represented channels. Each block contains convolutions that output 64, 128, 256 and 512 channels, respectively.

The final stage comprises of average pooling and a fully-connected layer to reduce dimensionality to the desired output (scalar for regression tasks, logits for classification tasks). The final block returns an activation tensor of shape $512\times6\times7\times6$. This tensor is flattened by fully average-pooling (C.-Y. Lee, Gallagher, and Tu 2016) the three spatial dimensions. A final fully-connected layer connects the 512 activations to the output, the dimensions of which are determined by the task.

3D convolutions are used throughout the network to leverage the full structure of the MRI. Convolutional operators in three-dimensions are trivially derived from one or two dimensions. Recall the 2D convolution operation from Chapter 2:

$$S_{i,j,k} = \sum_{l,m,n} I_{i+m-1,j+n-1,l} \cdot K_{k,m,n,l}$$

where K is the kernel of size $m \times n$ with output channels k operating on image I of size $i\times j$ with channels l. In the case of three spatial dimensions, this becomes:

$$S_{h,i,j,k} = \sum_{l,m,n,o} I_{h+o-1,i+m-1,j+n-1,l} \cdot K(k,m,n,o,l)$$

where h, i, j are the spatial dimensions of the image and m, n, o is the size of the kernel.

Three-dimensional convolutions carry a significant computational cost compared 2D convolutions. However, the advantage is that spatial information in all dimensions is kept. This is opposed to operating on independent 2D slices as that disassociates information in the dimension that is sliced across.

Kernel sizes for deep learning have become increasingly standardised with most models using an equilateral size of 3. Here, $3 \times 3 \times 3$ kernels with stride 2 are used throughout the network except for the very first layer, where a $7 \times 7 \times 7$ is used to

capture a larger region. A stride of 2 downsamples the image by a factor of 2 at each layer.

Deeper networks have more ability to learn representations and have shown empirically to perform better in most tasks (Tan and Le 2019; He et al. 2016). This is particularly true for networks with skip connections (e.g. residual connections) that are able to avoid vanishing gradients in very deep networks. However, the improved performance of deeper networks comes at increased computational and memory usage. Given the working resource constraint of a single GPU with 8 GB of memory, a network with 18-layers (shown in Figure 3.1) provided an appropriate balance between performance and resource cost. Deeper and shallower versions of the 3D-ResNet can also be constructed by adding or removing layers in the second stage.

I implemented the model using the python libraries PyTorch (Paszke, Gross, Chintala, et al. 2017) and TensorLy (Kossaifi, Panagakis, et al. 2019). Model development and prototyping was run on an NVIDIA GTX 1080 GPU. Models used for experiments were trained on NVIDIA P100 and RTX6000.

3.4 Experiments

To test the model, I set up four supervised learning tasks. Model input was T1-weighted brain structural MRI and the objective to predict a biological phenotype.

Biological sex. This binary classification serves as a demonstration task for model prototyping. The label distribution is approximately even with few missing labels. Anatomic differences between brain structures of males and females have been previously described (Kaczkurkin, Raznahan, and Satterthwaite 2019). No differences in function or ability are hypothesised or implicated otherwise.

Age. Unlike the previous task, this has potential medical applications. Brain structure changes significantly with age (Sowell, Thompson, et al. 1999; Jernigan et al. 1991) and is affected by a range of different traits and exposures (Raz et al. 2005). In Chapter 4 we will explore these in greater detail. Previous studies (Lemaitre et al. 2012) have used scalar measures of grey matter volumes (as a proxy for brain neuronal volume) to predict age from MRI. More recent studies have used CNNs to achieve significantly lower error rates.

Body mass index. The third task was to estimate Body Mass Index (BMI). Higher BMI is associated with increased brain atrophy (Ward et al. 2005) and lower cognitive function (Walther et al. 2010). This is a task that tests a model's ability to detect changes to brain that are influenced by both genetic predisposition and various exposures (Barness, Opitz, and Gilbert-Barness 2007).

Systolic blood pressure. The effects of Systolic Blood Pressure (SBP) on retinal images have been previously described (Poplin, Varadarajan, et al. 2018) in the context of predictive deep learning models. The ability of models to predict SBP from structural brain images remains poorly understood, but can have implications for risk stratification.

Cognition. Here, cognition is represented by an individual's score on the fluid intelligence test in the UK Biobank cognitive measurement battery. Brain structure is known to associate with cognitive decline (Tomasi and Volkow 2012).

3.4.1 Data visualisation set-up

We assume that the data lies on a high-dimension manifold. Estimating and visualising this manifold enables us to qualitatively evaluate the structure of the data and the inherent difficulty of classifying or regressing samples. For learning this manifold, I use 736 IDPs that are provided as part of UK Biobank's official imaging release. These phenotypes include various attributes describing global structure, such as volume of brain regions, intracellular volume fraction, orientation dispersion and isotropic volume fraction. After excluding individuals who did not have IDP records, the data contained 19 379 samples.

I reduce the dimensions with a manifold learning method: Uniform Manifold Approximation and Projection (UMAP) (McInnes and Healy 2018). UMAP is a graph representation learning algorithm that optimises a low-dimensional manifold to approximate a higher dimensional graph representation of the input data. For optimisation, I set the nearest neighbors to 200, minimum distance to 0.1 and minimise the Manhattan distance (L^1 norm). Visualisation results are presented in section 3.5.1.

3.4.2 Model training set-up

All regression tasks were trained end-to-end by minimising the MSE loss:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

where Y_i is the true label of the ith sample, $\hat{Y}_i$ the predicted label value and N is the number of samples. The objective of the classification task was to minimise the cross entropy:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N} Y_i \log(\hat{Y}_i)$$

Unless otherwise stated, this was optimised using Adam optimisation (Kingma and Ba 2015) for 50 epochs and mini-batch size of 8 with an initial learning rate of 10^{-4}, decaying by a factor of 10 at epochs 25, 50 and 75. To reduce overfitting and improve generalisation to unseen data I applied an L_2 weight decay of 5×10^{-4} on all parameters during training. I selected these hyperparameters by finding the combination that provided the minimum mean squared error on the validation set via a grid search. For all experiments I split the samples at random into train, validation and test sets. The three sets contained 11 520, 3 847 and 3 847 samples, respectively.

Constant baseline. To measure whether the models are learning complex relationships, I defined a simple baseline that uses only the *training labels* to make predictions. For binary classification tasks, this simple baseline predicts the majority (maximum occurrence) class of the training set, $\hat{y} = \arg\max_y |\{y \in \mathbf{y}_{train}\}|$. For regression tasks it predicts the mean of the training set, $\hat{y} = \bar{\mathbf{y}}_{train}$. If a model outperforms this baseline then it can be presumed to make use of information from the features to make improved predictions. In essence, this baseline serves as a lower bound for performance.

Plain 3D CNN baseline. An advanced baseline was also used for comparison, with *plain* signifying lack of residual connections. This 3D CNN model follows architecture proposed by Cole, Poudel, et al. (2017). The neural network has 11 trainable layers in repeated blocks of convolution $\rightarrow$ ReLU $\rightarrow$ convolution $\rightarrow$ batch norm $\rightarrow$ ReLU $\rightarrow$ max pooling. The final block is followed by a fully-connected layer that output a single value used for age prediction. The original network was optimised using SGD with momentum (Sutskever, Martens, et al. 2013). However,

I found that Adam optimisation performed better and gives a more direct comparison between the models, as opposed to comparing both model and optimiser simultaneously.

3.4.3 MRI data pre-processing

Used here are UK Biobank MRI scans[1] that have been minimally processed (Datafields 20252 and 20253 for T1 and T2-weighted scans, respectively), as described in section 3.2. Normalisation was performed on the input data to facilitate learning. For most of the MRI scans, voxel intensities (the 'brightness' of individual voxels) are not standardised between images. This can cause model training to become unstable due to discrepancies between back-propagation errors and previous model parameters. For example, a gradient update that is significantly smaller than the average will not change the model parameter values and updates that are too large can cause numerical overflow. To avoid this, I normalised and scaled the data to remove inter-image biases; each image was scaled to zero mean and unit standard deviation. In addition, I found that linearly scaling the target variables so that the regression training targets were $\in [0, 1]$ resulted in more stable training behaviour. Without this the model would output predictions only in a limited range that was smaller than the range of the target distribution.

A very small number of sample files were corrupt and unusable. These unusable files were marked accordingly by UK Biobank and excluded from any analysis or aggregate scores.

A small number of images had incorrect size. To alleviate this I cropped/padded irregular images to fit the most common set of dimensions. Padding was done using *edge* padding to expand the background canvas. Using `resize_image_with_crop_or_pad()`, a function included in the DLTK package (Pawlowski, Ktena, et al. 2017).

3.5 Results

The results of the data and task visualisations are presented first to give context for the difficulty of each task. Following this are the results of the 3D-ResNets trained on those tasks, before reporting the networks' sensitivity to hyperparameters.

[1] UK Biobank released scans with different levels of processing.

3.5.1 Task visualisation

The unsupervised manifold learning of the IDPs resulted in the structure presented in figure 3.2. Qualitative analysis reveals three large and connected clusters. The level of "connectedness" is highly dependent on the `minimum distance` UMAP parameter, which is set to 0.1 in these visualisations.

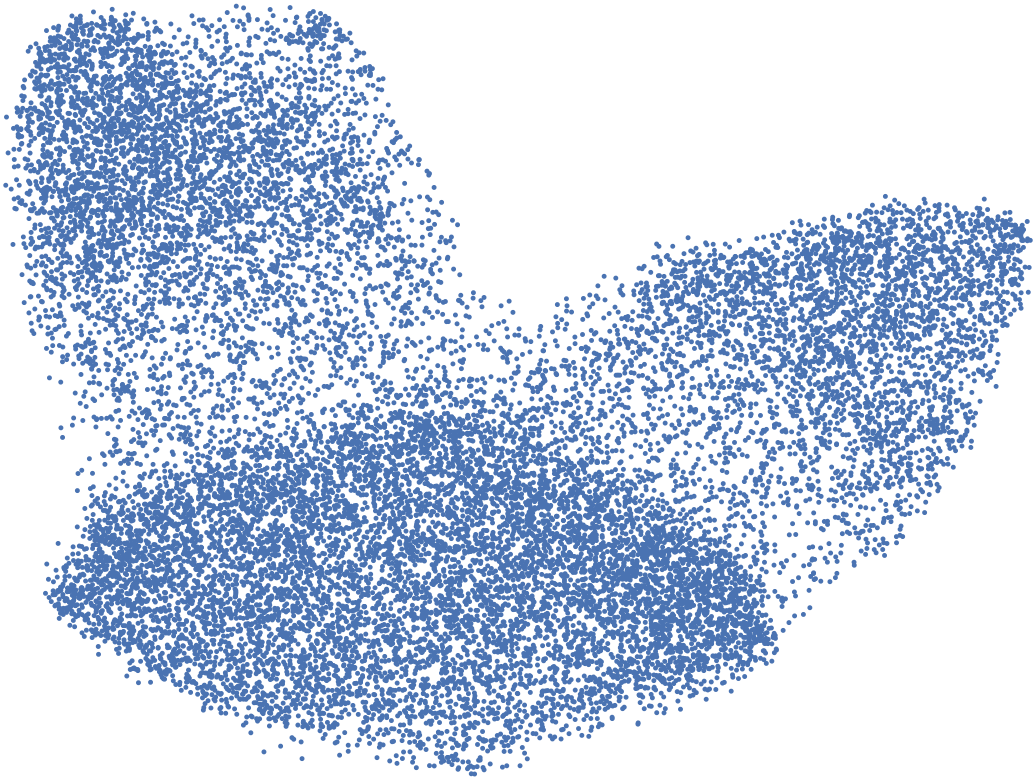


Figure 3.2: Unsupervised visualisation (with UMAP) of brain IDPs. Each point represents a single individual. The space of 736 IDPs was fitted onto a manifold in an unsupervised manner leading to clustering.

The relationships encoded by the structure are revealed by labelling the individual samples by trait in figure 3.3. Age is clearly embedded in the learned manifold with a trajectory from the top left cluster down through the lower central one, with older individuals embedded on the lower right. A more localised structure is observed when labelled by biological sex. Males and females are embedded within each of the three main clusters. This is expected as the global structured appears driven by age and the distribution of biological sex has been controlled for with respect to age in the cohort.

Labelling by fluid intelligence score unveils a weak global structure, with lower scores

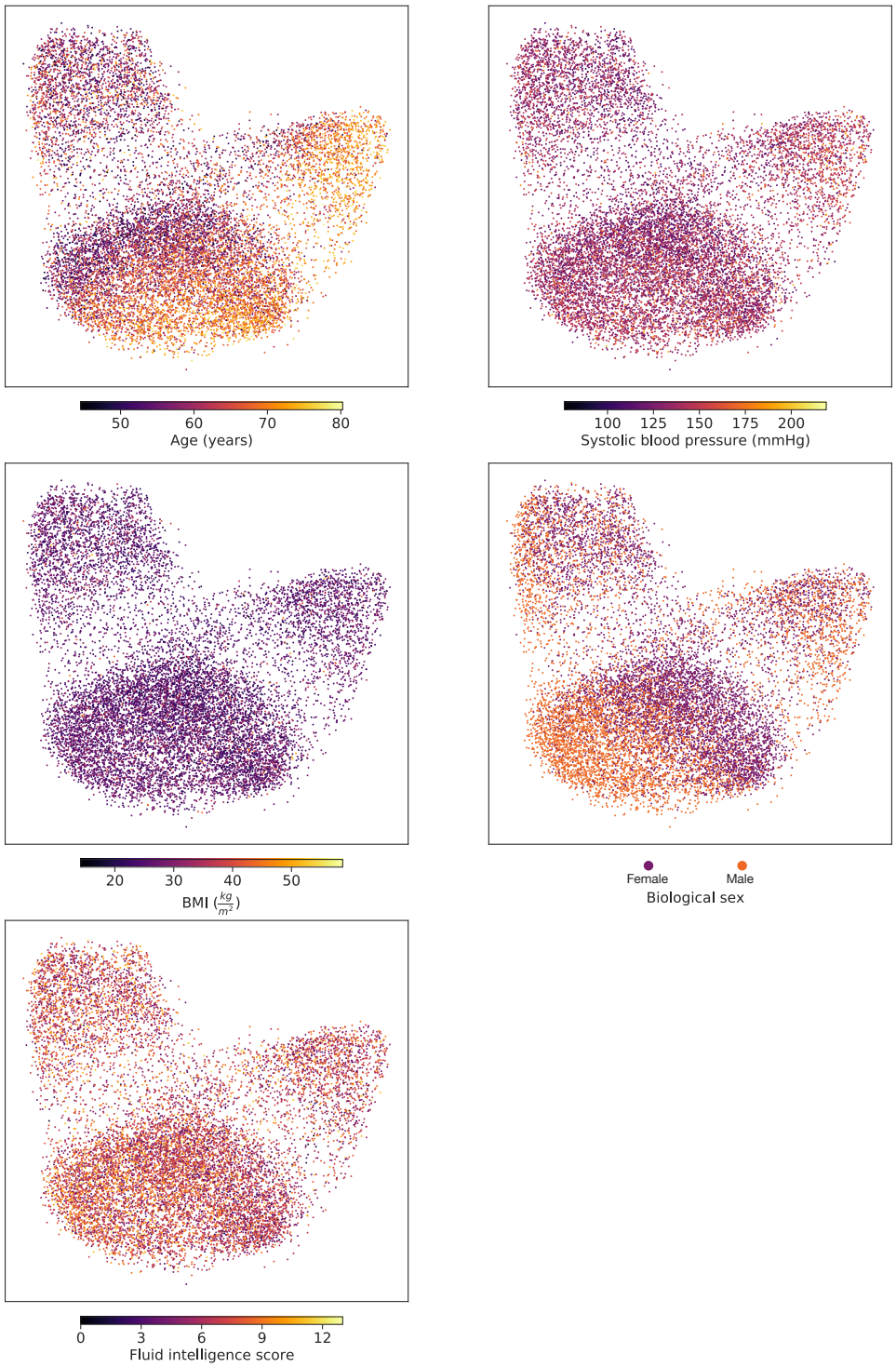


Figure 3.3: UMAP of brain IDPs, labelled with individual traits. Age shows strong global structure with older individuals embedded to the lower right of the manifold. Biological sex shows strong local structure with males and females embedded into sub-clusters within age groups. BMI, SBP and fluid intelligence did not exhibit obvious clustering.

in the bottom right of the manifold. This is consistent with the age structure and a relationship between age and cognitive decline has been well documented (Deary et al. 2009; Jurado and Rosselli 2007). No obvious structure is observed when labelled by either SBP or BMI. This suggests that the relationship between these two traits and brain structure represented by MRI could be weaker than between the MRI and age or biological sex.

3.5.2 Learning tasks with 3D-ResNets

The results of the 3D-Resnet are presented in table 3.2 along with comparisons to two baselines: a constant baseline that has access to only the training label distribution and a plain 3D CNN used in previous work for age prediction on MRI.

Learning models (i.e. plain 3D CNN and 3D-ResNet configurations) were able to outperform the constant baseline for classification of individuals by biological sex and prediction of BMI and age. This demonstrates the ability to learn useful relationships between input features and output label. The purpose of the constant baseline is to define a lower-bound for performance that learning models must outperform if learning useful representations.

The 3D-ResNets significantly outperform the plain 3D CNN on biological sex prediction, with a classification error of only 0.79 % compared with 6.70 % for plain 3D CNNs. These are both notably lower than the constant baseline, at an error of 45.82 %. Out of the three 3D-ResNet configurations tested (10-layer, 18-layer and 50-layer), the 50-layer 3D-ResNet performed best. Deeper networks could not be tested due to memory constraints.

For age prediction, arguably the most medically-relevant task, the 3D-ResNets again performed best; achieving an MAE on the test set of 2.47 years, with the deepest 3D-ResNet reporting the lowest error. Trained networks showed a small bias with true age. Similar bias has been reported and investigated in previous studies (S. M. Smith, Vidaurre, et al. 2019). I corrected for this by fitting a linear adjustment to the predictions on the training set. This increased the mean squared error but was done to keep the bias constant across age ranges[2].

BMI prediction is a more challenging task, as demonstrated by the plain CNN's failure to learn and 3D-ResNet-10's minor improvements over the baseline. The 18 and 50-layer 3D-ResNets were able to learn a much better representation, as

[2]Further investigation of this phenomena is presented in Chapter 4

demonstrated by their lower error rates. This is likely due to sufficient capacity to capture the relationship, which we know, from the unsupervised manifold learning, to be more complex than age prediction.

SBP prediction was difficult task. The 18 and 50-layer 3D-ResNets performed slightly better than the baselines. Further tests will are needed to check for statistical significance. Fluid intelligence prediction was not significantly better that the constant model for any of the CNNs tested.

Table 3.2: Performance of the models at learning four different tasks from brain MRI (UK Biobank test set). Deeper networks with more capacity perform better on most of the tasks. The constant baseline predicts the majority training class for binary classification tasks and as the mean training label for regression tasks. Fluid Intelligence Score (FIS) is the score from 0-13 representing the number of correct answers given in a fluid intelligence test.

Model	**Biological sex**	**Age**	**BMI**	**Systolic BP**	**FIS**
	Classification error	MAE	MAE	MAE	MAE
Constant baseline	45.82 %	7.06 years	3.37 kg/m^2	14.02 mmHg	1.67
Plain 3D CNN	6.70 %	4.85 years	4.37 kg/m^2	14.55 mmHg	3.32
3D ResNet-10	0.86 %	2.54 years	3.29 kg/m^2	13.25 mmHg	1.68
3D ResNet-18	0.79 %	2.68 years	2.14 kg/m^2	12.67 mmHg	1.67
3D ResNet-50	0.73 %	2.47 years	2.51 kg/m^2	12.82 mmHg	1.68

3.5.3 Hyperparameter analysis

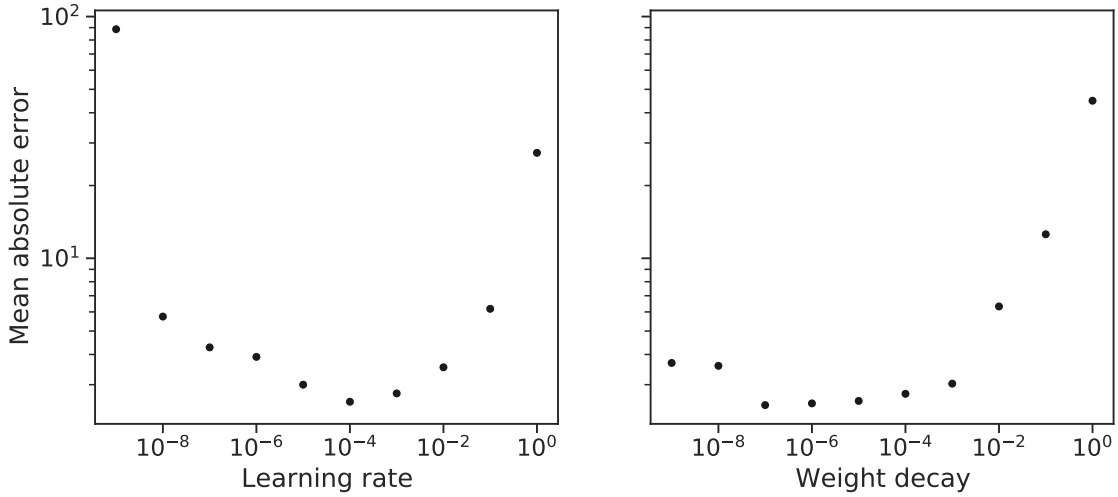


Figure 3.4: Effect of two hyperparameters on MAE of age prediction on the test set. Both initial learning rate (left) and weight decay (right) during training have substantial impact on the MAE.

In addition to comparing three different depth configurations of the 3D-ResNets across all the tasks, I performed detailed analysis of two hyperparameters: initial

learning rate and weight decay (L_2 penalty on all trainable weights). This was done to investigate the importance and sensitivity of different set-ups on the model's learning process for the task of age prediction. When varying the weight decay, the learning rate was set to 10^{-4} and when varying the learning rate, the weight decay was set to 5×10^{-4}. These are the same values as used for the results on medical tasks.

Results are shown in figure 3.4. The model is sensitive to the initial learning rate, with the MAE at the end of training being a logarithmically convex function of initial learning rate. The optimum initial learning is approximately 10^{-4} with significantly worse performance if starting above 10^{-2} or below 10^{-6}.

Model sensitivity to weight decay regularisation was also notable. Set-ups with weight decay below 10^{-2} gave good results, possibly due to higher values being too restrictive and dominating the total loss. Performance below 10^{-7} starts to tail off, probably a sign of overfitting.

3.6 Discussion

Unsupervised manifold learning of the data generated visible structure in the brain MRI IDPs for age, biological sex and fluid intelligence. These positive controls demonstrated the feasibility of the tasks to be learned with a supervised set-up. Visualisations also highlight the global and local structure of the data. Age embedding was visible on the global landscape with large separation between young and old individuals. Cognition, measured as the score on a fluid intelligence test, follows a similar manifold distribution to age, with regions dominated by younger individuals occupied by those with high fluid intelligence scores. The direct relationship between increased age and cognitive decline has been well documented (Deary et al. 2009; Jurado and Rosselli 2007). Surprisingly, age, but not fluid intelligence, could be learned by the models. The 3D-ResNet outperformed the baselines classification of by biological sex and prediction of BMI and age, with deeper networks performing better in general. Other tasks could not be learned by any of the models.

The structure of the data creates a set of opportunities and challenges. UK Biobank is a cohort study where features were not collected to investigate particular hypothesis. Instead, they provide a wide scope of various health-related attributes. This allows for both hypothesis-free analysis and, given prior knowledge, formed hypotheses to be tested. Two of the tasks I selected here: biological sex classification and age prediction, have been shown to be learnable (by e.g. Pawlowski and Glocker

(2019) and Cole, Poudel, et al. (2017), respectively) from brain structural MRI. Having this prior provided evidence that these tasks could be learned by a model with sufficient capacity and appropriate optimisation. Unsupervised manifold learning provided further justification that a relationship between the features and target exist. With no observed structure and limited prior work on some tasks (such as SBP), there was no assurance that they could be learnt from the data.

Having established that some of these tasks can be learnt, it is equally important to review *why* we want to learn them. Firstly, biological sex classification is done only to test the model and serves no applicable purpose. UK Biobank's selection resulted in an approximately even female:male ratio with the attribute recorded for almost every individual in the imaging cohort. This provided an evenly balanced and well-labelled set that served as a model demonstration. Predicting age from brain MRI is a very expensive and over-engineered solution to record someone's age. However, the age predictions themselves, more precisely the residuals of predictions, of a well-trained model carry important information. Deviations between predicted age and true are possibly compounded effects related to the genotype, environmental exposures and diseases that have affected an individual throughout their lifetime. These effects, their phenotypic associates and brain regions importance for age prediction are explored in detail in Chapter 4.

Similar analysis could potentially be used to analyse interactions between brain structure and either BMI or Systolic Blood Pressure (SBP). Chronic hypertension is associated with increased brain pathology (Muller et al. 2014) and obesity has been shown to correlate with changes in brain structure (Raji et al. 2010).

Another important aspect of this study is understanding *why* 3D-ResNets outperform the "plain" 2D neural network architectures. The two distinguishing aspects of 3D-ResNets are the residual (skip) connections and 3D convolutions. As discussed in the architectural overview, residual connections (He et al. 2016) enable much deeper networks to train much more effectively as they are more robust to vanishing gradients during error backpropagation. Intuitively, these connections allow a layer's input to "skip" over the layer's operations and be directed to the layer's output. Therefore, the layer's operation create deviation from the identity function. In the absence of skip-connections, a layer is learning deviations from a zero function (i.e. if the weights are set to zero, the layer's output will be zero). Furthermore, residual connections have favourable backward-pass properties as the scale of the error gradient is preserved throughout the backward-pass. This eliminates the vanishing gradient problem of non-residual networks. The properties

of these skip connections have been studied in detail by G. Yang and Schoenholz (2017). Note that they find that only α-ReLUs (whose activation is defined as $\psi_\alpha(x) = x_\alpha$ if $x > 0$ and 0 otherwise.) with $\alpha < 1$ have sub-exponential behaviour in the forward pass, therefore preventing exploding gradients. Contemporary work by Korolev et al. (2017) found that residual networks with 3D convolutions could be used for classifying individuals who have Alzheimer's disease.

Even though these models are capable of learning complex relationships, they are very sensitive to the set-up, namely hyperparameters, input pre-processing and optimisation. This is particularly true for end-to-end learning systems where feature selection and feature engineering has been replaced by architecture and hyperparameter tuning. Here, I investigated the effects of a few hyperparameters on the age prediction task. As expected, there were global minima for both the learning rate and weight decay that were then used for the final training setups. The learning rate can also be tuned by changing the learning rate schedule, which was fixed for all these experiments to reduce by a factor of 10 every 15 epochs. This was found to be optimal through heuristic testing as I did not have the computational resources to train with every combination of learning rate scheduling. For the same reasons, I do not account for interactions between hyperparameters which inevitably have significant correlations with each other. Finding the optimum (or near-optimum) in this high-dimensional hyperparameter is difficult. A random search would be a simple next step, and has been shown to work better than grid search in certain cases (Bergstra and Bengio 2012). A more systematic approach is to frame the hyperparameter selection as another task that can be learned directly. This is often called *neural architecture search* in the context of neural networks has in some cases found better models than those designed by humans (Zoph and Le 2017).

However, these setups are extremely computationally intensive and often only work with a limited set of operations and architectural sub-modules. Major advances are more likely to come from completely different models, rather than simply permuting a set of convolutional, activation function and normalisation operations. As an example of this, preliminary reports (Anonymous 2021) show that Transformers (Vaswani et al. 2017), which are traditionally used for natural language processing, perform exceptionally well on image classification.

In addition to model changes, the problem can be re-formulated. Approaching the task from a different perspective can make it easier for the models to learn. Given the problem setup: a large number of samples, a large number of multi-modal features and more than one target, a multi-task approach could provide many

benefits. Multi-task learning allows us to jointly model the data to learn a common representation, which can then be used to predict the various targets. A common representation with shared parameters would result in greater generalisability and robustness to noise. Tasks with fewer labelled samples available would also benefit because the model could leverage representations from other, better labelled tasks. This requires that the tasks are sufficiently related for the representations to the useful in other contexts. Attributes that are highly associated can in theory provide synergy. For example, age and cognitive tests scores are associated and therefore one would expect a model that predicts age well to be able to predict cognitive score not significantly worse than a linear model trained to predict cognitive score from age alone.

3.7 Conclusion

In this chapter I presented a neural network architecture that uses 3D spatial convolutions to predict biomedically-relevant phenotypes from raw brain structural MRI. The results demonstrate the importance of using the full 3D structure while remaining computationally efficient and practical to be run on large databases. Predicting phenotypes with accuracy is an important factor in risk stratification that can lead to improved health outcomes through better diagnosis and prognosis. These predictions can be either direct diagnoses (e.g. predicting cardiovascular disease from images) or risk stratification, where individuals can be rated for risk remediation. In the next chapter I will demonstrate how this system can be applied in the real-world to generate predictions that enhance the information available from medical imaging data.

Chapter 4

Structural brain changes as markers for disease

In this chapter we will look at an application of trait prediction from structural brain data. Specifically, how models that learn age from structural brain MRI can provide a useful measure for investigating systemic health, augment neuroradiological research and potentially serve as decision-support tools in clinical environments.

Some of the work presented in this section has been published as "Accelerated MRI-predicted brain ageing and its associations with cardiometabolic and brain disorders" (Kolbeinsson, S. Filippi, et al. 2020). I designed the study, developed methods, analysed the data, drafted and led in revision of the manuscript for intellectual content. Sarah Filippi designed study, and developed methods. Yannis Panagakis developed methods. Paul M Matthew supported design of the study and interpreted the data. Abbas Dehghan guided design of the study and interpreted the data. Joanna Tzoulaki jointly led in design of the study and interpreted the data. I use the plural "we" to acknowledge their contributions in this chapter.

4.1 Motivation

Chronological age is a major risk factor for poorer physical and mental health and chronic later life neurodegenerative diseases (Sowell, Peterson, et al. 2003; Stern 2012; Matthews et al. 2013). Brain structures and functions show considerable heterogeneity, suggesting that they change at different rates between individuals as a consequence of differences in genotype, environment or lifestyle and disease (Raz

et al. 2005). We therefore hypothesised that age-related differences between brains relative to changes in a "healthy" normative population may provide an index of disease or disease risk.

A variety of approaches have been used for multi-dimensional modelling of "brain age" from brain MRI images and for assessing associations and differences between modelled brain ages and specific health outcomes, exposures or traits (Cole, Marioni, et al. 2019). However, most prior studies have had relatively small sample sizes and have been applied in populations selected for a specific clinical pathology or outcome, as large-scale MRI phenotyping of large general populations has not been performed until recently (Sudlow et al. 2015b). Previous work also has relied on linear methods that cannot capture non-linear relationships within the data or "black box" machine learning methods unable to provide information concerning which brain image features were predictive, limiting the interpretability of findings.

Here we will see an application of the models and tools developed in the previous chapter to address these limitations, by using a CNN with T1-weighted brain MRI data from 21 382 volunteers in the UK Biobank. The CNN model for predicting brain age was trained on brain images from sub-groups selected for their relative health. It is possible to assess the relative contributions of features from different brain regions to identify those that were most informative for the model using permutation importance. The clinical relevance of differences between modelled and chronological brain age differences is then explored by associating these differences with over 1 400 clinical, lifestyle and environmental characteristics for individuals in a different group of 1 296 of UK Biobank volunteers that had not been stratified for relative health.

Definition (Brain age difference)**.** *Brain age difference is defined as the difference (measured in years) between an individual's chronological age and the age predicted by a model from brain features (here, T1-weighted brain MRI). For example, a 55-year-old individual whose age is predicted to be 47 years by the model has a brain age difference of -8 years.*

4.2 Methods

Here we have analysed the interim release of T1-weighted structural brain MRI on 21 382 participants from the UK Biobank imaging sub-study (K. L. Miller et al. 2016). The images were captured on a 3T Siemens Skyra scanner (software platform VD13). Each image was single channel with dimensions $182 \times 218 \times 182$ at $1\,\text{mm}^3$

resolution. Available reconstructed images were aligned to the MNI152 template (Jenkinson et al. 2002).

4.2.1 Study design

To set up a supervised machine learning framework, we split the data from 21 382 individuals into four sets: a training (N=3 067), validation ($N = 3\,962$), healthy test (N=2 057) and unselected (general) population test set (N=12 296). Individuals were divided randomly between the four sets, subject to restrictions on the number of apparently healthy and unhealthy individuals in each set. Healthy participants were defined as those who had no diagnoses in their hospital episode records available through UK Biobank, at the time of, or preceding, baseline assessment. The training set was used to optimise parameters of the neural network and learn relationships between healthy brain structure and age. This allowed the model to learn a representation of the physiological age-related modifications in brains of healthy individuals as a proxy for changes occurring with normal aging processes. Including brains from unhealthy individuals would have forced the model to learn structural changes related to disease and correct its predictions accordingly. The validation set was used to tune and assess the model without exposing it to individuals reserved for the final analysis.

I used two separate hold-out test sets to evaluate model performance. The former containing only healthy individuals to test the model on unseen healthy individuals. The latter test set was used for conducting the phenome-wide association study and included individuals who had not been stratified for health status. See Table 4.1 for descriptive characteristics of each data set.

Table 4.1: Distribution of healthy and unhealthy individuals defined by their diagnosis status at baseline (healthy: having no recorded diagnosis in hospital episode statistics data in UK Biobank) in the four data sets.

Dataset	**Healthy**	**Unhealthy**	**Total**
Train	3 067	0	3 067
Validation	1 071	2 891	3 962
Test - healthy	2 057	0	2 057
Test - unstratified	1 041	11 255	12 296

4.2.2 Deep neural network for brain age difference modelling

The objective was to determine the brain age difference for an individual, given an MRI of their brain and their chronological age. We defined brain age difference as the difference between an individual's chronological age and the age predicted by the deep learning model from T1-weighted MRI.

The age prediction model was a 3D-ResNet as described in Chapter 3 and took as input a T1-weighted structural image with dimensions $182 \times 218 \times 182$ and output the predicted age of the individual. The network was trained by minimising the mean squared error between the true and predicted values and optimized using Adam (Kingma and Ba 2015). We used 3D convolutions throughout the network to leverage the full structure of the MRI. Models were implemented in PyTorch (Paszke, Gross, Chintala, et al. 2017) and TensorLy (Kossaifi, Panagakis, et al. 2019), and trained on an NVIDIA P100 GPU.

4.2.3 Interpretability estimation with permutation importance

Understanding how deep neural networks use the features available to it is not straightforward. The networks' sequences of non-linear mappings are extremely unintuitive for humans. In this analysis, we used a permutation importance approach to analyse the importance of different brain regions by quantifying their contribution to model predictions. The method was proposed as a way of interpreting random forest models (Breiman 2001). It works by repeatedly permuting a specific feature, in this case a region of the brain, between individuals. Serial repetition across the population develops a distribution of predictions without the feature, which can be compared with that before data removal.

The brain regions were defined in the UK Biobank MNI152 atlas (available as part of UK Biobank data). For each brain region, the approach consisted of permuting the images corresponding to that region between individuals (e.g. the right insular cortex of individual A gets moved to individual B, while the same region of individual B gets moved to individual C and so on), predicting the ages of all individuals based on these modified brain MRI and measuring the accuracy of our original model to predict age for all individuals with the switched region. We quantified the loss of accuracy in terms of increased mean absolute error, the same metric we used to train the unperturbed model. For a region that was important for age prediction,

a degradation in performance and increased prediction error would result. Conversely, permuting regions that contributed insignificantly to the prediction would result in a smaller drop in accuracy.

4.2.4 Statistical analysis

I performed a phenome-wide association study to test for associations between brain age difference and 1 410 phenotypic characteristics measured on UK Biobank participants through clinical assessments, record linkage and health and lifestyle questionnaires. This agnostic scan was performed using the PHEASANT analysis method as previously described in the literature (Millard et al. 2017). The software uses a series of regression analyses (linear for continuous traits and logistic for binary traits) to associate traits with the exposure of interest (in this case, brain age difference). For each trait, samples with missing values were excluded from that analysis. All analyses were adjusted for age, sex and assessment centre. To account for multiple testing, we used Bonferroni correction ($P = 2.35 \times 10^{-5}$). In further sensitivity analyses, to account for the correlation between measured phenotypes in UK Biobank, we used a false discovery rate (FDR) of 5% using the Benjamini–Hochberg procedure to account for multiple testing (Benjamini and Hochberg 1995).

4.2.5 Mendelian randomisation

Causality of associations between brain age differences and traits is difficult to infer as the data were collected cross-sectionally as part of an observational study. Mendelian randomisation provides a method for assessing the causal nature of some associations by using genetic variants as instrumental variables for risk factors (Davey Smith and Ebrahim 2003). We performed two-sample Mendelian randomisation analysis of selected traits that had summary GWAS data available (diastolic and systolic blood pressure, pulse pressure, Alzheimer's disease and diabetes) to explore the causality of the reported associations. Genetic variants used as instruments were obtained from DIAGRAM 1000G study for type II diabetes (Morris et al. 2012), from the International Genomics of Alzheimer's Project (Lambert et al. 2013) for Alzheimer's diseases, and from a recent study (Evangelou et al. 2018) for blood pressure. For the latter, we used the allele effects from the International Consortium for Blood Pressure (Ehret et al. 2011) to avoid bias due to overlapping samples. For the association of the genetic variants with brain age difference, we used the data from UK Biobank as available on October 17$^{\text{th}}$, 2019. We matched

and harmonised the effective allele for each set of instruments with brain age difference and removed the correlated variants using linkage disequilibrium (LD) clumping ($r^2 < 0.1$). We estimated the causal effects using the inverse variance weighted method. Potential pleiotropic effect was detected using heterogeneity tests and sensitivity analysis was done using weighted median and MR Egger regression methods to rule out pleiotropic effects. All analyses were done using the Two Sample MR package (Hemani et al. 2018).

4.3 Results

4.3.1 Brain age prediction using the optimised neural network

The model-defined and chronological ages were strongly correlated in the training set of healthy individuals ($N = 3\,067, \text{Pearson correlation} = 0.97, \text{MAE} = 1.71\,\text{years}$). However, the neural network showed a linear bias for age; individuals older/younger than the cohort average were predicted to be younger/older than they are. Similar bias has been reported and investigated in previous studies (S. M. Smith, Vidaurre, et al. 2019). After linearly adjusting the model output for chronological age using data from the training set, the model showed no significant bias on the validation set ($N = 3\,926$).

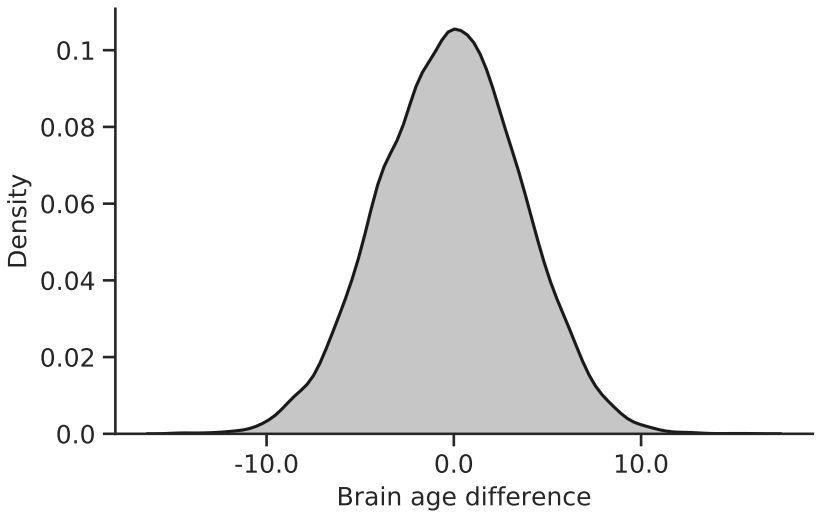


Figure 4.1: Distribution of brain age differences across the test cohort. Standard deviation is 3.72 years.

I then applied the model to the larger test set of individuals who had not been stratified for health ($N = 12\,196$). The model achieved a mean absolute error of 3.42 years in a unimodal distribution of differences between predicted and chronological ages (Figure 4.1). There was a strong direct relationship (Figure 4.2, Pearson correlation $= 0.82, P = 2.67 \times 10^{-242}$). We separately explored accuracy on the test set containing only healthy individuals ($N = 2\,057$) which had a lower mean absolute error of 2.87 years.

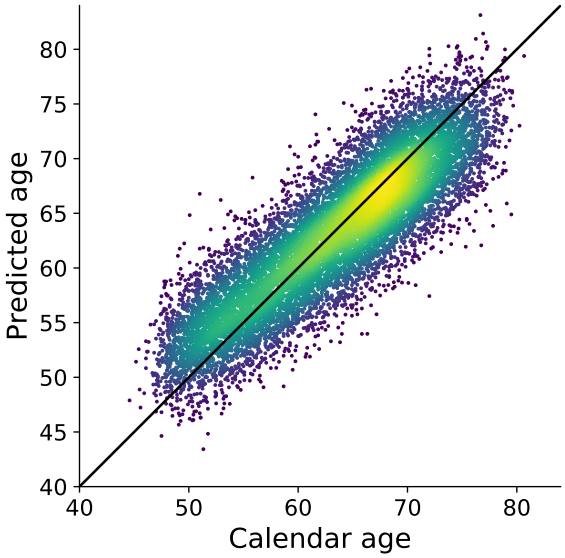


Figure 4.2: Age predicted by the deep neural network developed here, and linearly adjusted for age using coefficients calculated from the training set, plotted against calendar age for all participants in the test set. The diagonal line is $y = x$, or a perfect predictor. Colour indicates the density of the scatter with brighter being denser. The Pearson correlation is 0.82.

4.3.2 Contributions of different brain regions to brain age predictions

I attempted to partially explain those image features contributing most to the age prediction model. We assessed differential contributions of brain regions at the level of major white and grey matter regions by serial inference of the model with image regions permuted between individuals. Six brain regions (left cerebellar lobules I-IV and the left crus and vermis, the right hippocampus, left amygdala, and left insular cortex) were found to contribute most to the accuracy of age prediction (Figures 4.3 and 4.4). Removing the information contributions of any of those regions caused the mean absolute error to increase by more than 0.10 years.

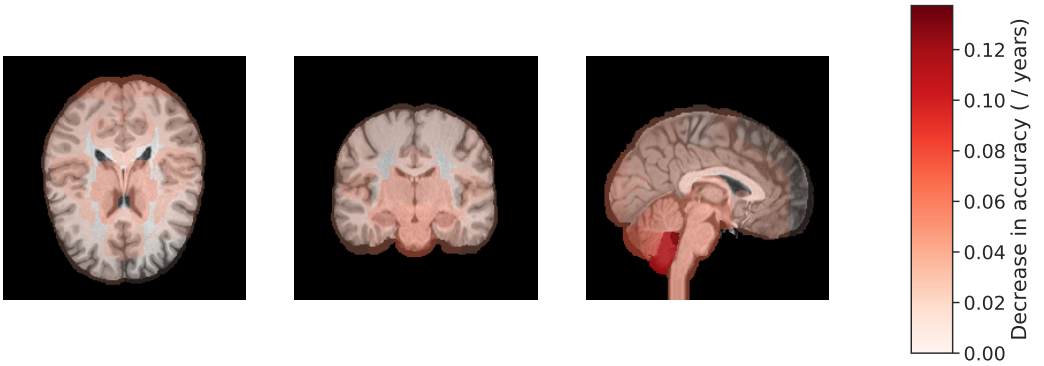


Figure 4.3: Regions of the brain highlighted by importance on age predictions from T1-weighted brain MRI. Brighter colour overlay indicates more important regions as defined by permutation importance.

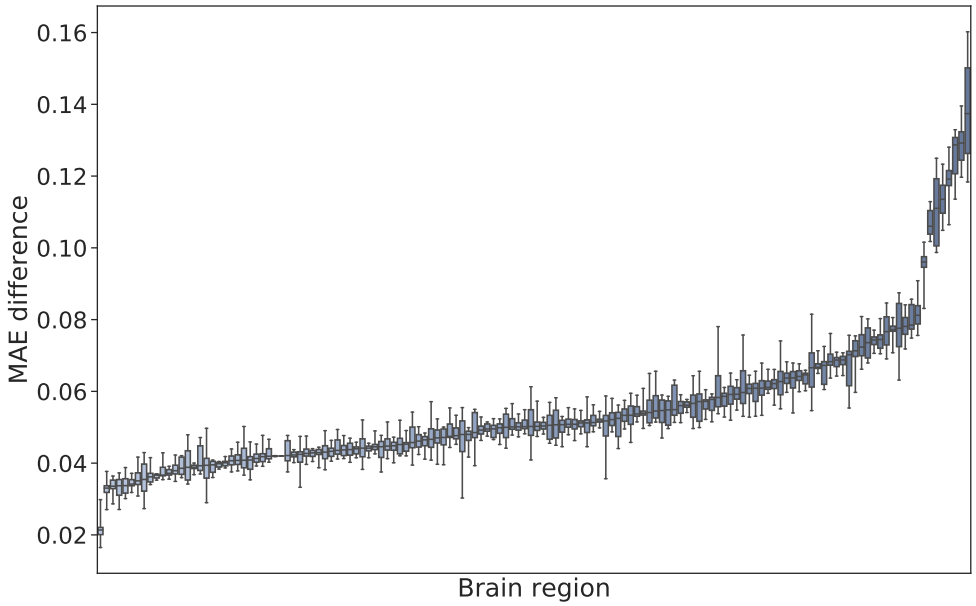


Figure 4.4: Impact on each of the 140 brain regions (ordered along x-axis) on MAE of age prediction accuracy (y-axis). This was calculated using permutation importance, see section 4.2.3. It illustrates the effect of removing information contained within the region of interest by replacing it with voxels from the same region in another, random sample, run over the entire test set ($N = 12\,296$). The whiskers are the range of samples obtained for 12 repeated trials.

4.3.3 Phenome-wide association study

I then explored the potential meaningfulness of differences between model predicted and chronological ages in the unstratified test population. We evaluated associations of these differences with more than 1 410 of the International Statistical Classification of Diseases and Related Health Problems (ICD) codes, self-reported clinical conditions and physical, lifestyle and environmental phenotypes. Of these, 24 were found to be significantly associated with brain age differences after correcting for multiple testing ($P < 2.35 \times 10^{-5}$) (20 were direct associations, 4 inverse associations). These results are visualised in Figure 4.5 and listed in Table 4.2.

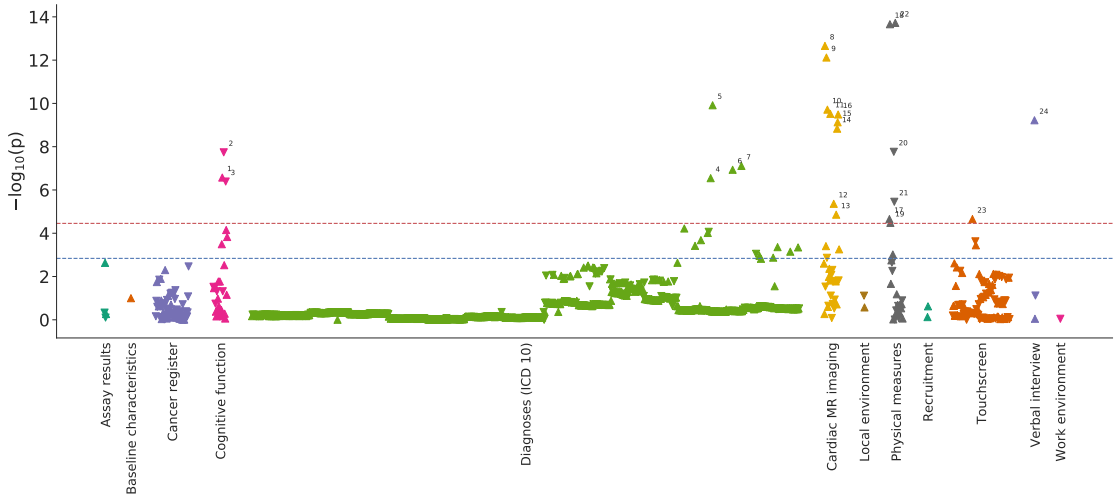


Figure 4.5: Manhattan type plot showing the significance of association (p-value) between 1 410 UK Biobank traits and brain age difference, coloured by trait category. The Bonferroni-corrected significance threshold is marked by a horizontal red line (p-value = 2.35×10^{-5}) and the 5 % FDR correction threshold with a blue line (p-value = 1.45×10^{-3}). Trait label 1: Time taken to start entering values in symbol-digit matching test, 2: Number of symbol digit matches made correctly, 3: Number of symbol digit matches attempted, 4: Multiple sclerosis, 5: Essential (primary) hypertension, 6: Diagnoses - secondary ICD10: Type 1 diabetes, 7: Type 2 diabetes, 8: Systolic brachial blood pressure during pulse wave analysis (PWA), 9: Central systolic blood pressure during PWA, 10: Cardiac index during PWA, 11: End systolic pressure during PWA, 12: Stroke volume during PWA, 13: Central augmentation pressure during PWA, 14: Cardiac output during PWA, 15: Central pulse pressure during PWA, 16: Peripheral pulse pressure during PWA, 17: Ventricular rate, 18: Diastolic blood pressure, 19: Body mass index (BMI), 20: Hand grip strength (left), 21: Hand grip strength (right), 22: Systolic blood pressure, 23: Taking insulin, 24: Number of treatments/medications taken.

The diagnoses and traits associated with brain age differences included cardiovascular and metabolic diseases and risk factors, cognitive function and physical strength.

Table 4.2: Traits that associate with brain age difference, with p-value ¡ 2.35×10^{-5} (Bonferroni threshold) . Odds ratios and betas are given per unit standard deviation of brain age difference (3.72 years). PWA = pulse wave analysis.

Categorical and ordered traits	**Category**	**Odds ratio (95% CI)**	**p-value**	**Rate of incidence (cases/controls)**
Multiple sclerosis	Diagnoses	4.04 (2.37, 6.93)	2.85×10^{-7}	18/12278
Type 1 diabetes	Diagnoses	2.39 (1.73, 3.29)	1.17×10^{-7}	49/12247
Taking insulin	Touchscreen	2.22 (1.55, 3.14)	2.22×10^{-5}	36/1633
Type 2 diabetes	Diagnoses	1.42 (1.25, 1.61)	7.77×10^{-8}	334/11962
Essential (primary) hypertension	Diagnoses	1.22 (1.15, 1.29)	1.22×10^{-10}	1820/10476
Number of treatments/medications taken	Verbal interview	1.13 (1.09, 1.18)	6.02×10^{-10}	12294
Continuous traits	**Category**	**Beta (95% CI)**	**p-value**	**N samples**
Systolic brachial blood pressure during PWA	Heart MRI	0.08 (0.06, 0.10)	2.20×10^{-13}	10338
Diastolic blood pressure	Physical measures	0.08 (0.06, 0.10)	2.20×10^{-14}	11830
Central systolic blood pressure during PWA	Heart MRI	0.08 (0.06, 0.10)	7.59×10^{-13}	10337
Systolic blood pressure	Physical measures	0.07 (0.06, 0.09)	1.91×10^{-14}	11830
End systolic pressure during PWA	Heart MRI	0.07 (0.05, 0.09)	3.04×10^{-10}	10306
Peripheral pulse pressure during PWA	Heart MRI	0.07 (0.05, 0.09)	3.25×10^{-10}	10335
Time taken to enter values in symbol-digit matching test	Cognitive function	0.07 (0.04, 0.09)	2.67×10^{-7}	6592
Central pulse pressure during PWA	Heart MRI	0.07 (0.04, 0.09)	7.47×10^{-10}	10335
Cardiac output during PWA	Heart MRI	0.07 (0.04, 0.09)	1.47×10^{-9}	10188
Cardiac index during PWA	Heart MRI	0.07 (0.05, 0.09)	1.94×10^{-10}	10046
Stroke volume during PWA	Heart MRI	0.05 (0.03, 0.07)	4.36×10^{-6}	10190
Ventricular rate	Physical measures	0.05 (0.03, 0.07)	2.23×10^{-5}	10656
Central augmentation pressure during PWA	Heart MRI	0.05 (0.03, 0.07)	1.39×10^{-5}	10333
Body mass index	Physical measures	0.04 (0.02, 0.06)	3.33×10^{-5}	12260
Hand grip strength (right)	Physical measures	-0.03 (-0.05, -0.02)	3.51×10^{-6}	12267
Hand grip strength (left)	Physical measures	-0.04 (-0.05, -0.03)	1.72×10^{-8}	6581
Number of symbol digit matches attempted	Cognitive function	-0.07 (-0.09, -0.04)	4.08×10^{-7}	6581
Number of symbol digit matches made correctly	Cognitive function	-0.07 (-0.10, -0.05)	1.82×10^{-8}	10338

For example, there was a four-fold higher risk of having been diagnosed with multiple sclerosis (OR 4.04, 95 % CI 2.38–6.93) for each positive SD difference between predicted and chronological age (3.7 years). Measures of blood pressure also showed positive associations with brain age difference including the self-reported diagnosis of hypertension (OR 1.22, 95 % CI 1.15–1.29 per SD increase in brain age difference) and measured systolic (beta 0.07, 95 % CI 0.06–0.09) and diastolic (beta 0.08, 95 % CI 0.06–0.10) blood pressures. Direct associations were found between brain age difference and metabolic traits such as type I (OR 2.39, 95 % CI 1.73–3.29) and type II (OR 1.42, 95 % CI 1.25–1.61) diabetes and participants taking insulin (OR 2.22, 95 % CI 1.55–3.14) (Table 1).

Conversely, individuals with a predicted age younger than their chronological age were found to have greater physical strength, as reflected in hand grip strength (beta -0.03, 95 % CI -0.05 to -0.02). Brain age differences were inversely associated with improved performance in tests included in the UK Biobank cognitive battery, including time taken to enter values in a digit-symbol matching test (beta 0.07, 95 % CI 0.04-0.09) and the numbers of symbols matched correctly (beta -0.07, 95 % CI -0.10 to -0.05).

With a less stringent threshold of 5 % FDR ($P = 1.45 \times 10^{-3}$), 20 additional associations with brain age difference were observed (16 direct associations, four inverse associations, Table 4.3). Additional positive associations with brain age difference included having had a depressive episode (OR 3.33, 95 % CI 1.85 to 6.01), history of a prior psychiatric episode (OR 1.97, 95 % CI 1.35 to 2.87) and higher neuroticism score (OR 1.07, 95 % CI 1.03 to 1.12).

Mendelian randomization analysis of selected traits

I used Mendelian randomisation to investigate the effects of genetic determinants for a range of traits on brain age differences in order to explore the potential for causality (Supplementary Table 3). The association of a higher genetically determined diastolic blood pressure with higher brain age difference in main and sensitivity regression analyses (inverse variance weighted: beta 0.06, p-value 0.01, weighted median: beta 0.07, p-value 0.02, MR Egger: beta 0.14, p-value 0.05) (4.4) provided evidence in support of a causal relationship of blood pressure. The MR-Egger intercept, a measure of pleiotropy that may bias the main inverse variance weighted estimate, did not reach statistical significance, further supporting this conclusion. There was also no evidence for heterogeneity (Cochrane's Q for inverse variant weighted median: 0.14, Cochrane's Q for MR Egger: 0.15). By contrast, there was

Table 4.3: Traits that associate with brain age difference, with 2.35×10^{-5} (Bonferroni threshold) <p-value <1.45×10^{-3} (FDR threshold). Odds ratios and betas are given per unit standard deviation of brain age difference (3.72 years).

Categorical and ordered traits	**Category**	**Odds ratio (95% CI)**	**p-value**	**Rate of incidence (case/control)**
Depressive episode	Diagnoses	3.33 (1.85, 6.01)	6.04×10^{-5}	15/12281
Occupational therapy and vocational rehabilitation	Diagnoses	3.27 (1.75, 6.16)	2.09×10^{-4}	13/12283
Diabetic retinopathy	Diagnoses	2.57 (1.44, 4.59)	1.34×10^{-3}	15/12281
Diagnoses - secondary ICD10: Z50.1 Other physical therapy	Diagnoses	2.02 (1.37, 3)	4.39×10^{-4}	33/12263
One or more previous psychiatric episodes with this Health Care Provider	Psychiatric	1.97 (1.35, 2.87)	3.83×10^{-4}	35/10988
Gastro-intestinal haemorrhage	Diagnoses	1.89 (1.29, 2.76)	1.11×10^{-3}	35/12261
Calculus of kidney	Diagnoses	1.64 (1.25, 2.17)	4.52×10^{-4}	67/12229
Epilepsy	Diagnoses	1.63 (1.23, 2.17)	7.12×10^{-4}	64/12232
Personal history of long-term (current) use of anticoagulants	Diagnoses	1.43 (1.2, 1.72)	9.72×10^{-5}	162/12134
Total errors traversing alphanumeric path (trail #2)	Trail making	1.17 (1.08, 1.28)	3.22×10^{-4}	2331
Neuroticism score	Psychosocial factors	1.07 (1.03, 1.12)	3.61×10^{-4}	10355
On hormone replacement therapy	Health and medical history	0.81 (0.72, 0.91)	2.37×10^{-4}	583/1087
Single live birth	Diagnoses	0.76 (0.67, 0.87)	8.58×10^{-5}	546/11750
Second degree perineal laceration during delivery	Diagnoses	0.64 (0.49, 0.83)	8.91×10^{-4}	106/12190
Continuous traits	**Category**	**Beta (95% CI)**	**p-value**	**N sample**
Visceral adipose tissue volume	Abdominal MRI	0.06 (0.03, 0.09)	3.94×10^{-4}	3627
Duration to complete alphanumeric path (trail #2)	Touchscreen	0.06 (0.03, 0.08)	7.12×10^{-5}	5802
Duration to complete numeric path (trail #1)	Touchscreen	0.06 (0.03, 0.08)	1.49×10^{-4}	5802
Heart rate during PWA	Heart MRI	0.04 (0.02, 0.06)	5.55×10^{-4}	10416
Waist circumference	Physical measures	0.03 (0.01, 0.05)	9.58×10^{-4}	12287
LV stroke volume	Heart MRI	-0.06 (-0.10, -0.02)	1.36×10^{-3}	2857

no evidence for a causal influence of diabetes on brain age differences.

Table 4.4: Mendelian randomisation results.

Exposure	**Method**	**Beta**	**p-value**	**Intercept p-value**	**Heterogeneity p-value**
Alzheimer's disease	Inverse variant weighted	-0.02	0.36	N/A	0.83
	Weighted mean	-0.01	0.67	N/A	N/A
	MR Egger	-0.02	0.60	0.95	0.80
Diabetes	Inverse variant weighted	0.03	0.57	N/A	0.64
	Weighted mean	0.10	0.21	N/A	N/A
	MR Egger	0.18	0.18	0.23	0.66
Diastolic blood pressure	Inverse variant weighted	0.06	0.01	N/A	0.14
	Weighted mean	0.07	0.02	N/A	N/A
	MR Egger	0.14	0.05	0.23	0.15
Systolic blood pressure	Inverse variant weighted	0.01	0.22	N/A	0.08
	Weighted mean	0.03	0.08	N/A	N/A
	MR Egger	0.05	0.34	0.54	0.07
Pulse pressure	Inverse variant weighted	-0.03	0.04	N/A	0.003
	Weighted mean	-0.04	0.28	N/A	N/A
	MR Egger	-0.17	0.317	0.37	0.003

4.4 Discussion

In this large study of 21 382 middle and older aged participants with rich brain MRI data, we developed a deep learning approach to calculate brain age difference with respect to a healthy reference population. Brain age difference should reflect cumulative effects on brain structure associated with effects of environmental, lifestyle and disease exposures, as well as individual differences in genotype. We have approached the "explainability" of this measure by characterising the brain regions whose features made the greatest contributions to brain age difference, which were discovered to be the cerebellum, hippocampus, amygdala and insular cortex. Finally, we conducted an exploration of over 1 400 phenotypes and traits and demonstrated associations between brain age difference and clinically meaningful traits related to cardiovascular, metabolic and brain diseases.

Previous studies have used measures of difference or changes in brain or grey matter volumes over time to provide indirect measures of the relative rates of brain neuronal volume loss during life (Lemaitre et al. 2012). However, relationships between brain structures and relative signal intensities (e.g., those reflected in MRI

"texture" measures) also change (Kovalev, Kruggel, and Cramon 2003). Unlike simple scalar volume measures, these structural and tissue image texture changes are highly multi-dimensional. This has led to the use of multivariate non-linear machine learning methods such as neural networks, which use high-dimensional MRI features to predict age-related changes in the brain (Cole, Poudel, et al. 2017). Prior studies using these have shown associations between deviations of predicted ages from chronological ages (similar to the brain age difference metric described here) and mortality (Cole, Ritchie, et al. 2018), obesity (Kolenic et al. 2018), malnutrition (Franke, Gaser, Roseboom, et al. 2018), brain trauma (Cole, Leech, et al. 2015) and psychiatric disorders (Kaufmann et al. 2019). Analogues to this approach have been described in other contexts, e.g., relating to the increased shortening of telomeres with respect to age in a population with multiple sclerosis (Krysko et al. 2019).

This study used a large population and applied an advanced deep learning approach that results in more accurate predictions of age from MRI than has been reported in previous efforts (Cole, Poudel, et al. 2017; Cole, Ritchie, et al. 2018; Kolenic et al. 2018; Gaser et al. 2013). To distinguish the influences of clinical diagnoses and traits, the neural network is implicitly constrained to learn a representation of the concept from a cohort selected for being healthy. The differences between predicted and chronological ages therefore captures changes in brain structure related to disease or disease risk relatively independently from those related to normal healthy biological ageing.

The brain regions identified in the interpretation analysis play central roles in cognition and memory (hippocampus), emotional regulation and salience (amygdala) and physiological homeostasis (insula). All of these regions have been recognised previously as having a functional role or showing population differences relevant to brain health (Gunning-Dixon et al. 2003; Hartley et al. 2014; Menon and Uddin 2010). The importance of the cerebellum for brain age is of particular interest, as relationships between cerebellar pathology and cognitive dysfunction or late life neurodegenerative diseases remain poorly described (Jacobs et al. 2017; Hoche et al. 2018). This emphasises the importance of applying agnostic learning models which can be coupled with analysis for describing specific features with greater granularity. Their discovery could contribute to better understanding of the mechanisms underlying relationships between brain structure and health.

Brain structural changes, as captured by the index, correlated with clinical diagnoses and phenotypic characteristics, as well as cognitive function, extending prior

studies based on different models and using different training sets that also described deviations of predicted ages from chronological ages (similar to the brain age difference metric described here) amongst pathological or "at risk" subgroups of the larger cohort (Bashyam et al. 2020; Ning et al. 2020; S. M. Smith, L. T. Elliott, et al. 2020). Here, individuals with higher brain age difference performed worse in cognitive tests for fluid intelligence, giving support to the index as an informative metric. In terms of diseases, participants with multiple sclerosis had higher brain age differences, revealing changes to brain in addition to those that are age-related. Indeed, multiple sclerosis is associated with macro- and microscopic inflammatory and demyelinating pathology in both white and grey matter (Filippi et al. 2000) and has previously been associated with increased brain age differences (Cole, Raffel, et al. 2020; Høgestøl et al. 2019). In contrast, we did not observe a statistically significant association between brain age difference and cerebral infarction. This is possibly due to cerebral infarction causing heterogenic brain changes between individuals that are not consistent enough to be associated with a measure across a population. Alternatively, our stringent multiple testing corrections to guard against false positives can potentially cause some true positives to be missed; a limitation of the approach. Analysis showed associations between higher brain age difference and type I and type II diabetes, a finding previously observed in other imaging studies (Franke, Gaser, Manor, et al. 2013) including analysis on UK Biobank 10 and supported by previously recognised effects of diabetes on brain structure (Kodl and Seaquist 2008; Suzuki et al. 2019). This study also revealed a direct association between brain age differences and vascular disease risk factors, particularly blood pressure. Associations between brain age deviations using distinct approaches to ours but similar data from UK Biobank also revealed associations between this phenotype and blood pressure, adding internal validity to these results (Cole 2020). Although a relationship between hypertension and both cognitive decline and brain atrophy has already been established (Suzuki et al. 2019), and prolonged hypertension is recognised to be associated with increased white matter pathology (De Leeuw et al. 2002), the mechanisms of these associations are not well defined (Qiu and Fratiglioni 2015). Physical fitness was associated with lower brain age difference highlighting the association of physical fitness not only to functional (Angevaren et al. 2008), but also to structural changes of the brain.

We examined the potential for causality in the associations between traits and brain age differences. MR analysis demonstrated a likely causal relationship between increased diastolic blood pressure and increased brain age. This implies that reducing

diastolic blood pressure would have an impact on the relative brain age, broadly consistent with clinical evidence that reducing or preventing hypertension reduces the risk of strokes (Howard et al. 2015). By contrast, our MR analysis did not provide evidence for a causal effect of diabetes on brain age differences, suggesting common, pleotropic factors (pleiotropy) may contribute causally to both, consistent with the possibility that treatments for diabetes mellitus also may have an independent impact on late-life neurodegenerative processes (Meng et al. 2020). Future prospective studies relating brain age differences and the incidence of cognitive impairments would add to confidence that the measure could be used as a risk stratification tool for late life cognitive impairments or other brain disorders.

Although adopted here, the term "brain age" should be used cautiously. Aging is referenced to time since birth, but incorporates concepts of time-dependent intrinsic biological processes and individually specific influences acting on a tissue or person (Rowe and Kahn 1987); changes to the brain during the life course are more than solely a consequence of time (chronological age) alone. The use of a healthy population for training is intended to maximise the interpretability of brain age difference as an index of risk of dysfunction or disease.

As noted above, the sensitivity of the approach to factors affecting individuals is limited, as the interpretability analysis is based on population characteristics. Further work could focus on developing descriptions and explanatory hypotheses at an individual level, e.g. using methods such as Shapely Additive Explanations (Lundberg and S.-I. Lee 2017).

Another important limitation is the accuracy of region definitions. Although the images were carefully aligned as part of UK Biobank pre-processing, imperfect alignment would cause boundary effects. These were partially mitigated by running the permutation multiple times and averaging the results. However, the relative accuracy of detection of disease or disease risk associations depends on the population sample size and structure; our detections of associations are, for example, impacted by the relatively low prevalence of stroke, multiple sclerosis and diabetes in the population studied. We are making our model openly available for others but need to highlight that it was developed and validated with UK Biobank data; generalisation would require extending training to include new target populations or data acquired using different MRI platforms or sequences.

We adopted a hypothesis free approach to investigate a range of phenotypes in relation to brain age differences with adjustment for multiple comparisons. However,

different associations had different sample sizes and therefore power to detect associations which should be taken into account when interpreting the associations.

4.5 Conclusion

In this large study of 21 382 middle and older aged participants with rich brain MRI data, we developed a deep learning approach to calculate brain age difference. Brain age difference should reflect cumulative effects on brain structure associated with effects of environmental, lifestyle and disease exposures, as well as individual differences in genotype. We have approached the "explainability" of this measure by characterising the brain regions whose features made the greatest contributions to brain age difference, which were discovered to be the cerebellum, hippocampus, amygdala and insular cortex. Finally, we conducted an exploration of over 1 400 phenotypes and traits and demonstrated associations between brain age difference and clinically meaningful traits related to cardiovascular, metabolic and brain diseases.

These results add to a growing literature demonstrating the use of brain structural differences as a general marker of systemic health. They suggest that brain age difference may be an index of health or risk of later life metabolic, cardiovascular and brain diseases and functional traits relevant to health. Consistent with conclusions from large cohort treatment studies (Howard et al. 2015), our results suggest a direct causal link between higher diastolic blood pressure and brain age difference. With larger populations and further advances to learning methods and analysis, the approach may help to better define risk factors of brain disease. Stratifying people on this form of index may help identify individuals who could benefit most from interventions for brain health risk factor reduction. Finally, this work highlights the great potential for using machine learning as a decision-support tool to enhance the information available from neuroradiological reporting.

Chapter 5

Tensor learning

The central challenge of many machine learning tasks, both supervised and unsupervised, is to disentangle relevant information from noise and other signals irrelevant to the task. Tensor decompositions provide a way of separating the data, represented as tensors, into smaller factor tensors. Constraints can be applied to these factor tensors to reduce the computational complexity, build in inductive biases and facilitate learning (Kolda and Bader 2009). A system of tensor decompositions can be thought of as a *tensor network* (Cichocki et al. 2016), a term I will use to describe the addition of tensor decomposition layers into deep neural networks.

In this chapter I introduce two methods that incorporate tensor decompositions within deep learning. First, tensor dropout: a technique to improves generalisability and robustness to perturbations, both random noise and adversarial (section 5.2). In essence, tensor dropout works by dropping, at random, some ranks of the factor tensors during *composition*. I demonstrate the improvements of tensor dropout on both natural image and brain MRI tasks. The other method I introduce is a high-order Kronecker machine for multi-modal learning (section 5.3). By combining representations of multiple sources in a low-rank Kronecker space, it becomes tractable to model the sources' higher-order interactions. I demonstrate the feasibility of this method on integrating T1-weighted and T2-weighted brain MRI scans.

5.1 Tensor essentials

Here I will overview some of the basic and common notation that will be used in sections 5.2 and 5.3.

A *scalar* a is a zero dimensional object, a *vector* $\mathbf{v}$ is a set of elements ordered along a single dimension. A *matrix* $\mathbf{m}$ is an object with elements arranged in two dimensions. We can extend this idea into N dimensions, resulting in an object $\mathcal{X}$ of order N, which we term *tensor*. A tensor is a general concept that can represent structured data. However, they are naturally apt at representing Euclidean data. Note that these are sometimes specified as *data tensors* in the literature to separate them from the multilinear transformations that map between sets of vector spaces and from stress tensors in engineering.

5.1.1 Tensor operations

The Hadamard product of two matrices $\mathbf{A} \in \mathbb{R}^{M \times N}$ and $\mathbf{B} \in \mathbb{R}^{M \times N}$ is defined as the matrix:

$$\mathbf{A} * \mathbf{B} = \begin{bmatrix} a_{11}b_{11} & \dots & a_{1N}b_{1N} \\ \vdots & \ddots & \vdots \\ a_{M1}b_{M1} & \dots & a_{MN}b_{MN} \end{bmatrix}$$

which is also of size $M \times N$. Demonstrated here in the matrix case, this can be generalised to higher orders. For tensors of dimension $\mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, the resulting tensor, $\mathcal{C}$, will be of the same shape with element $c_{i_1,i_2,\dots,i_N} = a_{i_1,i_2,\dots,i_N} b_{i_1,i_2,\dots,i_N}$

The *Kronecker product* of two matrices $\mathbf{A} \in \mathbb{R}^{M \times N}$ and $\mathbf{B} \in \mathbb{R}^{P \times Q}$ is defined as the matrix:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1N}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{M1}\mathbf{B} & \dots & a_{MN}\mathbf{B} \end{bmatrix}$$

which is of size $MP \times NQ$.

The *Khatri Rao product* of $\mathbf{A} \in \mathbb{R}^{M \times N}$ and $\mathbf{B} \in \mathbb{R}^{P \times N}$ is a matrix

$$\mathbf{A} \odot \mathbf{B} = \begin{bmatrix} \mathbf{a_1} \otimes \mathbf{b_1} & \dots & \mathbf{a_N} \otimes \mathbf{b_N} \end{bmatrix}$$

which is of size $MP \times N$. Here, $\mathbf{a_i}$ and $\mathbf{b_i}$ represent the i$^{\text{th}}$ column of matrix $\mathbf{A}$ and $\mathbf{B}$, respectively. The Khatri Rao product can be described as the columnwise Kronecker product. It is particularly useful when operating on tensors where the first dimension is independent, e.g. representing samples or mini-batches. In those

cases we do not want to multiply elements between samples.

5.1.2 Tensor decompositions

With the basic operations described above, we can now define decompositions that break a tensor down into smaller factors.

Canonical Polyadic (CP) decomposition (Hitchcock 1927). Also known as tensor rank decomposition, PARAFAC, and CANDECOMP. This method can be thought of as an extension of Singular Value Decomposition (SVD) into higher dimensions. A tensor $\mathcal{X}$ is typically only approximated with a decomposition, only in rare cases is the original full tensor reproduced exactly. The CP decomposition for a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is given as:

$$\mathcal{X} \approx \sum_{r=1}^{R} \mathbf{I}_{1,r} \circ \mathbf{I}_{2,r} \circ \ldots\ldots \mathbf{I}_{N,r}$$

where R is the rank of the decomposition and defines the rank of the factors, $\mathbf{I}_{i,r}$ is the i$^{\text{th}}$ rank-1 tensor and $\circ$ is the outer product.

Tucker decomposition (Tucker 1966) breaks a tensor $\mathcal{X}$ down into a smaller core tensor $\mathcal{G}$ and a set of factor matrices, one for each dimension. $\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_N}$ where r_n is the chosen rank of dimension n. The corresponding factor-matrix along dimension n is $\mathbf{U}^{(n)} \in \mathbb{R}^{r_n \times i_n}$. The decomposition is written as:

$$\mathcal{X} \approx \mathcal{G} \times_0 \mathbf{U}^{(0)} \times_1 \mathbf{U}^{(1)} \times \cdots \times_N \mathbf{U}^{(N)}$$

Tensor Train (Ivan V Oseledets 2011) for $\mathcal{X} \in \mathbb{R}^{i_1 \times i_2 \times \cdots \times i_N}$ can be described as

$$\mathcal{X} \approx \sum_{\alpha=1}^{R} \mathbf{U}^{(1)}(i_1, \alpha)\mathbf{U}^{(2)}(i_2, \alpha) \ldots \mathbf{U}^{(d)}(i_d, \alpha)$$

For a comprehensive list and description of tensor methods and decompositions, please refer to the monograph *Tensor decompositions and application* by Kolda and Bader (2009).

5.1.3 Other tensor methods

From the base tensor operations and decompositions, a large number of composite tensor methods have been introduced. I will describe one such methods as a preliminary for the following section.

Tensor Regression Layer (TRL) (Kossaifi, Lipton, et al. 2020) combines low-rank decompositions with deep neural networks' weight tensors for parameter reduction. By defining a weight tensor as a composition of these low-rank factors, only the factors have to be learned. This leads to a significantly reduced parameter space that retains much of the expressiveness of the full tensor. I will use this tensor layer as a testbed to implement and test the tensor dropout method. The TRL is defined here, with a rank–$(R_1, \cdots, R_N)$ Tucker decomposition but without loss of generality. For a deep neural network with input (typically from the previous layer) $\mathcal{X}^{(k)}$, weight tensor $\mathcal{W}$, bias b, the output $y^{(k)}$ will be given as:

$$\begin{aligned} y^{(k)} &= \langle \mathcal{X}^{(k)}, \mathcal{W} \rangle + b \\ \text{with } \mathcal{W} &= \mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)} \end{aligned} \tag{5.1}$$

with $\mathcal{G} \in \mathbb{R}^{R_1 \times \cdots \times R_N}$, $\mathbf{U}^{(k)} \in \mathbb{R}^{I_k \times R_k}$ for each k in $[1 \,..\, N]$, $\mathbf{U}^{(N)} \in \mathbb{R}^{1 \times R_N}$ and O is the output dimensionality.

5.2 Tensor dropout

Some of the work presented in this section has been published as "Tensor Dropout for Robust learning" (Kolbeinsson, Kossaifi, et al. 2021) © IEEE 2021. This was a collaborative project. Jean Kossaifi (NVIDIA AI) and Yannis Panagakis (University of Athens) contributed to the idea conceptualisation, experiment guidance, results interpretation, and drafting and editing of the manuscript. Adrian Bulat (Samsung AI) performed experiments that are presented in the paper manuscript but omitted here. Anima Anandkumar (Caltech) contributed to the conceptualisation of the model and interpretation of the results. Ioanna Tzoulaki (Imperial College London) and Paul Matthews (Imperial College London) contributed to the interpretation of UK Biobank results. I use the plural "we" to acknowledge their contributions in this section.

5.2.1 Introduction

The overparameterisation of deep neural networks is a double-edged sword. On one side its capacity allows for powerful predictive models with remarkable performance on computer vision tasks (Krizhevsky, Sutskever, and Geoffrey E Hinton 2012; LeCun, Bengio, and G. Hinton 2015; He et al. 2016). On the other side it makes them notoriously prone to overfitting (Caruana, Lawrence, and Giles 2000). In addition, deep networks are not robust to small perturbations to the input; changes imperceivable to humans can lead to arbitrarily different predictions by the network (I. J. Goodfellow, Shlens, and Szegedy 2015). Not only does this significantly degrade trust in these networks, they are also chronically over-confident when predicting on noisy inputs (Hein, Andriushchenko, and Bitterwolf 2019). These weaknesses hinder the deployment of such systems in medical settings, and other fields that are sensitive to trust reliability.

Adding regularisation to these networks has been shown to improve robustness to adversarial perturbations (Bietti et al. 2019; Jakubovitz and Giryes 2018). Regularisation techniques can be classed as either: methods that constrain the parameters directly or structural changes to the architecture to make them inherently more robust. In the context of parameter constraints, regularisation can be applied in the form of added randomness to the activations (e.g. dropout (Srivastava et al. 2014)) or to the weights (e.g. DropConnect (Wan et al. 2013)). Regularisation functions (e.g. ℓ_1- or ℓ_2- norm) can also be directly applied to network's parameters (Nowlan and Geoffrey E. Hinton 1992; Krogh and Hertz 1991; Scardapane et al. 2017; Yuchen Zhang, J. D. Lee, and Jordan 2016).

However, the second class of methods: structural changes to the network itself, is arguably preferable in many cases. A prime example of such structural inductive biases are tensor methods that account for the structure in the data. Tensor methods allow us to fully leverage the structure in that data as the transformation from the input to the output can be generalised as a tensor map. Preserving multi-dimensional structure is crucial for performance. By limiting the network to matrix operations (e.g. with flattening layers followed by one or more fully-connected layers), we are ignoring this structure, resulting in deteriorated performance (Kossaifi, Khanna, et al. 2017; Kossaifi, Lipton, et al. 2020; Kossaifi, Bulat, Panagakis, et al. 2019). Tensor methods allow to leverage that structure to improve the model, reduce the number of parameters and improve computational efficiency. One way this is done is by leveraging multi-linear correlations in the network (Tai et al. 2016; Y. Cheng et al. 2015; Yu et al. 2017; Kossaifi, Lipton, et al. 2020; Kossaifi, Bulat,

Tzimiropoulos, et al. 2019).

This section introduces tensor dropout and details its application it to the TRL. Specifically, we propose a new stochastic rank-regularisation, applied to low-rank tensors in decomposed forms. This formulation is general and can be applied to any type of decomposition. We introduce it here, without loss of generality, to the case of Tucker and CP decompositions.

Summary of contributions:

- *Tensor dropout*, a novel stochastic tensor decomposition where non-linear dropout is applied in the latent subspace spanned by a low-rank factorisation.
- The application of tensor dropout to tensor regression layers and show that it improves the inductive bias of CNNs by fully leveraging the structure in the data via stochastic tensor decomposition.
- Demonstration of state-of-the-art performance for large scale regression from MRI data and that the model is significantly more robust to noise in the input, as occurs naturally during capture.
- Show that the method makes neural networks significantly more robust to adversarial noise, *without* adversarial training.
- Show that the method implicitly regularises the tensor decomposition. We establish theoretically and empirically the link between tensor dropout and the deterministic low-rank tensor regression.

5.2.2 Tensor dropout

I propose a novel randomised decomposition on the weight tensor $\mathcal{W}$, which applies dropout in the latent subspace spanned by a tensor decomposition. For instance, for an N^{th}–order regression weight with a Tucker structure, we can define for each $k \in [1 \,..\, N]$, a sketch matrix $\mathbf{M}^{(k)} \in \mathbb{R}^{R_k \times R_k}$ (e.g. a random projection or column selection matrix). This can then be used to sketch the factors $\mathbf{U}^{(k)}$ of the decomposition as $\tilde{\mathbf{U}}^{(k)} = \mathbf{U}^{(k)}(\mathbf{M}^{(k)})^\top$ and the core tensor, $\mathcal{G}$ is sketched as $\tilde{\mathcal{G}} = \mathcal{G} \times_1 \mathbf{M}^{(1)} \times \cdots \times_N \mathbf{M}^{(N)}$.

In the context of tensor regression, we can apply this tensor dropout technique to the weights. Given an activation tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ and a set of S target labels

$\mathbf{y}^{(k)}$ we can define a new type of layer, a Randomised Tensor Regression Layer (R-TRL) from equation 5.1, which uses Tensor Dropout and aims at minimising the empirical risk:

$$\frac{1}{S}\sum_{k=1}^{S}\left(\mathbf{y}^{(k)} - \langle\tilde{\mathcal{W}}, \mathcal{X}^{(k)}\rangle\right)^2, \tag{5.2}$$

where $\tilde{\mathcal{W}}$ is a stochastic low-rank approximation of Tucker decomposition. In other words, in addition to the low-rank structure of the weights, we apply our tensor dropout in the latent subspace spanned by the decomposition. For the proof, please refer to Kolbeinsson, Kossaifi, et al. (2021).

For instance, in the case of a Tucker R-TRL, we have:

$$\tilde{\mathcal{W}} = \tilde{\mathcal{G}} \times_0 \tilde{\mathbf{U}}^{(0)} \times \cdots \times_N \tilde{\mathbf{U}}^{(N)} \tag{5.3}$$

This is the core of our proposed R-TRL, which incorporates tensor dropout within a TRL. Even though several sketching methods have been proposed, the focus here is on R-TRL with two different types of binary sketching matrices, namely binary matrix sketching with replacement and binary diagonal matrix sketching with Bernoulli entries, which we detail below.

5.2.3 Bernoulli Tucker randomised tensor regression

For any $n \in [0 \,..\, N]$, let $\boldsymbol{\lambda}^{(n)} \in \mathbb{R}^{R_N}$ be a random vector, the entries of which are i.i.d. Bernoulli(θ), then a diagonal Bernoulli sketching matrix is defined as $\mathbf{M}^{(n)} = \text{diag}(\boldsymbol{\lambda}^{(n)})$.

When the low-rank structure on the weight tensor $\tilde{\mathcal{W}}$ of the TRL is imposed using a Tucker decomposition, the randomised Tucker approximation is expressed as:

$$\begin{aligned}\tilde{\mathcal{W}} =& \mathcal{G} \times_1 \mathbf{M}^{(1)} \times \cdots \times_N \mathbf{M}^{(N)} \\ &\times_0 \left(\mathbf{U}^{(1)}(\mathbf{M}^{(1)})^\top\right) \times \cdots \times_N \left(\mathbf{U}^{(N)}(\mathbf{M}^{(N)})^\top\right) \\ &= [\![\tilde{\mathcal{G}};\, \tilde{\mathbf{U}}^{(0)}, \cdots, \tilde{\mathbf{U}}^{(N)}]\!]\end{aligned} \tag{5.4}$$

The main advantage of considering the above-mentioned sampling matrices is that the products $\tilde{\mathbf{U}}^{(k)} = \mathbf{U}^{(k)}(\mathbf{M}^{(k)})^\top$ or $\tilde{\mathcal{G}} = \mathcal{G} \times_1 \mathbf{M}^{(1)} \times \cdots \times_N \mathbf{M}^{(N)}$ are never explicitly computed, the elements are selected from $\mathcal{G}$ and the corresponding factors.

Interestingly, in analogy to dropout, where each hidden unit is dropped indepen-

dently with probability $1-\theta$, in the proposed randomised tensor decomposition, the columns of the factor matrices and the corresponding fibres of the core tensor are dropped independently and consequently the *rank* of the tensor decomposition is stochastically dropped.

The tensor dropout acts as an implicit regulariser on the regression, by limiting the rank, at each iteration. This can be shown by examining the expectation of the stochastic loss, which can be expressed deterministically as the unrandomised empirical loss, plus an additional regularisation term.

Theorem 1. *Tensor Dropout with Tucker decomposition is a deterministic regularised loss.*

The minimisation objective (equation 5.2) can be reformulated by expanding the tensor contractions, and the expectation of the minimisation objective becomes:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\lambda}}\big[\frac{1}{S}\sum_{k=1}^{S}\Big(\mathbf{y}^{(k)}-\langle\tilde{\mathcal{W}},\mathcal{X}^{(k)}\rangle\Big)^2\big]\\
&=\frac{1}{S}\sum_{k=1}^{S}\big(\mathbf{y}^{(k)}-\theta^N\langle\mathcal{W},\mathcal{X}^{(k)}\rangle\big)^2\\
&+\frac{\theta^N(1-\theta^N)}{S}\sum_{k=1}^{S}\langle\mathcal{G}^{\star 2}\times_1(\mathbf{U}^{(1)})^{\star 2}\cdots\times_N(\mathbf{U}^{(N)})^{\star 2},(\mathcal{X}^{(k)})^{\star 2}\rangle.
\end{aligned}\tag{5.5}$$

5.2.4 Bernoulli CP randomised tensor regression

An interesting special case of equation (5.3) is when the weight tensor $\tilde{\mathcal{W}}$ of the TRL is expressed using a CP decomposition. In that case, we set $\mathbf{M}=\mathbf{M}^{(1)}=\cdots=\mathbf{M}^{(N)}=\text{diag}(\boldsymbol{\lambda})$, with, for any $k\in[1\,..\,R]$, $\lambda_k\sim\text{Bernoulli}(\theta)$.

Then a randomised CP approximation is expressed as:

$$\tilde{\mathcal{W}}=\sum_{k=1}^{R}\tilde{\mathbf{U}}_k^{(1)}\circ\cdots\circ\tilde{\mathbf{U}}_k^{(N)}\tag{5.6}$$

The above randomised CP decomposition on the weights is equivalent to the following formulation:

$$\tilde{\mathcal{W}}=[\![\boldsymbol{\lambda};\,\mathbf{U}^{(1)},\cdots,\mathbf{U}^{(N)}]\!]\tag{5.7}$$

Based on the previous stochastic regularisation, for an activation tensor $\mathcal{X}$ and a corresponding label vector $\mathbf{y}$, the optimization problem for our tensor regression layer with stochastic regularisation is given by:

$$\min_{\mathbf{U}^{(1)},\cdots,\mathbf{U}^{(N)}} \left(\mathbf{y}^{(k)} - \langle [\![\boldsymbol{\lambda}; \mathbf{U}^{(1)}, \cdots, \mathbf{U}^{(N)}]\!], \mathcal{X}^{(k)} \rangle\right)^2 \tag{5.8}$$

In addition, the above stochastic optimisation loss can be rewritten as a deterministic regularised one:

Theorem 2. *Tensor Dropout with CP decomposition is equivalent to a deterministic regularised loss.*

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\lambda}}\Big[\frac{1}{S-1}\sum_{k=1}^{S}\left(\mathbf{y}^{(k)} - \langle [\![\boldsymbol{\lambda}; \mathbf{U}^{(1)}, \cdots, \mathbf{U}^{(N)}]\!], \mathcal{X}^{(k)} \rangle\right)^2\Big] \\
&= \frac{1}{S}\sum_{k=1}^{S}\left(\mathbf{y}^{(k)} - \theta\langle [\![\mathbf{U}^{(1)}, \cdots, \mathbf{U}^{(N)}]\!], \mathcal{X}^{(k)} \rangle\right)^2 \\
&+ \frac{\theta(1-\theta)}{S}\sum_{k=1}^{S-1}\langle [\![(\mathbf{U}^{(1)})^{\star 2}, \cdots, (\mathbf{U}^{(N)})^{\star 2}]\!], (\mathcal{X}^{(k)})^{\star 2}\rangle.
\end{aligned} \tag{5.9}$$

5.2.5 R-TRL with replacement

Previously, we focus on the R-TRL with Bernoulli sampling only. Our model is more general and can be applied to many different sampling settings. Here, we introduce one such case: the R-TRL with a binary sketching matrix sampled with replacement. Specifically, we first choose $\theta \in [0, 1]$.

Mathematically, we then introduce the uniform sampling matrices $\mathbf{M}^{(1)} \in \mathbb{R}^{R_1 \times R_1}, \cdots, \mathbf{M}^{(N)} \in \mathbb{R}^{R_N \times R_N}$. $\mathbf{M}_j$ is a uniform sampling matrix, selecting K_j elements, where $K_j = R_j \text{ div } \theta$. In other words, for any $i \in [1 \,..\, N]$, $\mathbf{M}^{(i)}$ verifies:

$$\mathbf{M}^{(i)}(j,:) = \begin{cases} & \text{if } j > K \\ \mathbf{Id}_m(r,:), m \in [1 \,..\, R_i] & \text{otherwise} \end{cases} \tag{5.10}$$

In practice this is done efficiently by selecting directly the correct elements from $\mathcal{G}$ and its corresponding factors.

5.2.6 Experiments

In this section, we introduce the experimental setting, databases used and implementation details. We experimented on several datasets across various tasks, namely image classification and MRI-based regression. All methods were implemented using PyTorch (Paszke, Gross, Massa, et al. 2019) and TensorLy (Kossaifi, Panagakis, et al. 2019). For all adversarial attacks, Foolbox (Rauber, Brendel, and Bethge 2017) was used.

Phenotypic trait prediction from MRI data In the regression setting, we investigate the performance of our R-TRL in a challenging, real-life application on a very large-scale dataset. This case is particularly interesting since MRI volumes are large 3D tensors, all modes of which carry important information. The spatial information is traditionally discarded during the flattening process, which we avoid by using a tensor regression layer. In these experiments we train the entire model, but any pre-trained network can be easily modified post-hoc to make use of the TRL.

UK Biobank brain MRI dataset (Sudlow et al. 2015a) is the world's largest MR imaging database of its kind. The aim of the UK Biobank Imaging Study is to capture MRI scans of vital organs for 100 000 primarily healthy individuals by 2022. Associations between these images and lifestyle factors and health outcomes, both of which are already available in the UK Biobank, will enable researchers to improve diagnoses and treatments for numerous diseases. The data used here consists of T1-weighted $182 \times 218 \times 182$ MR images of the brain for 7 500 individuals captured on a 3 T Siemens Skyra system. 5 700 are used for training, 800 are used for validation and 1 000 samples are used to test. The target label is the age for each individual at the time of MRI capture. We use skull-stripped images that have been aligned to the MNI152 template (Jenkinson et al. 2002) for head-size normalization. We then center and scale each image to zero mean and unit variance for intensity normalisation.

Architecture	**Regression**	**MAE**
3D-ResNet	FC	3.23 years
3D-ResNet	Tucker	2.99 years
Ours	Randomized Tucker	**2.77 years**
Ours	Randomized CP	**2.58 years**

Table 5.1: Classification accuracy for UK Biobank MRI. The ResNet models with R-TRL significantly outperforms the version with a fully-connected (FC) layer.

Results: For MRI-based experiments we implement an 18-layer ResNet with three-dimensional convolutions. We minimize the mean squared error using Adam (Kingma and Ba 2015), starting with an initial learning rate of 10^{-4}, reduced by a factor of 10 at epochs 25, 50, and 75. we train for 100 epochs with a mini-batch size of 8 and a weight decay (L_2 penalty) of 5×10^{-4}. For Tucker-based R-TRL we used a tensor with rank $128 \times 6 \times 7 \times 6$. For CP-based R-TRL we used a Kruskal tensor with 82 components. As previously observed, our randomized tensor regression network outperforms the 3D-ResNet baseline by a large margin, Table 5.1. To put this into context, the current state-of-art for convolutional neural networks on age prediction from brain MRI on most datasets is an MAE of around 3.6 years (Cole, Poudel, et al. 2017).

Robustness study: I tested the robustness of our model to white Gaussian noise added to the MRI data. Noise in MRI data typically follows a Rician distribution but can be approximated by a Gaussian for signal-to-noise ratios (SNR) greater than 2 (Gudbjartsson and Patz 1995). As both the signal (MRI voxel intensities) and noise are zero-mean, we define SNR $= \frac{\sigma^2_{\text{signal}}}{\sigma^2_{\text{noise}}}$, where σ is the variance. We incrementally increase the added noise in the test set and compare the error rate of the models.

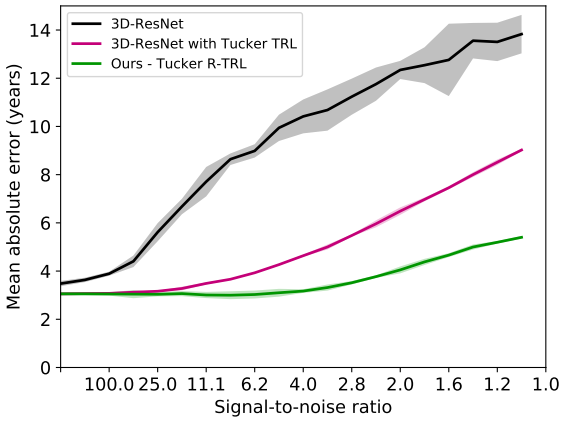
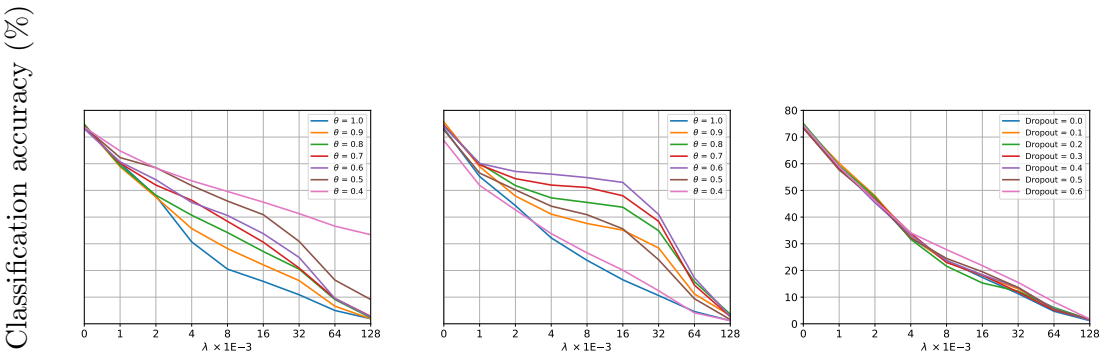


Figure 5.1: Age prediction error on the MRI test set as a function of increased added Gaussian noise. Shaded regions indicate 95% confidence intervals for 5 independent runs. A baseline model that predicts the average age from the training set would achieve an MAE of 7.6 years.

The ResNet with R-TRL is significantly more robust to added white Gaussian noise compared to the same architectures without it (figure 5.1). At signal-to-noise ratios below 10, the accuracy of a standard fully-connected ResNet is worse than a naive model that predicts the mean of training set (MAE = 7.9 years).



(a) FGS attack on Bernoulli Tucker R-TRL with different drop rates.

(b) FGS attack on CP R-TRL with different drop rates.

(c) FGS attack on Tucker TRL with different dropout rates.

Figure 5.2: Robustness to adversarial attacks, measured by adding adversarial noise to the test images, using the Fast Gradient Sign, on CIFAR-100 and Bernoulli drop. We compare a Tucker tensor regression layer with dropout applied to the regression weight tensor (Subfig. 5.2c) to our randomized TRL, both in the Tucker (Subfig. 5.2a) and CP (Subfig. 5.2b) case. Our approach is more robust.

Brain morphology is an important attribute that has been associated with various biological traits including cognitive function and overall health (Pfefferbaum et al. 1994; Swan et al. 1998). By keeping the structure of the brain represented in MRI in every layer of the architecture, the model has more information to learn a more accurate representation of the entire input. Randomly dropping the rank forces the representation to be robust to confounds. This is a particularly important property for MRI analysis since intensities and noise artifacts can vary significantly between MRI scanners (L. Wang et al. 1998). Randomized tensor regression layers enable both more accurate and more robust trait predictions from MRI that can consequently lead to more accurate disease diagnoses.

Ablation Studies on CIFAR-100 In the image classification setting, we perform a thorough study of this method on the CIFAR-100 dataset. We empirically compare this approach to both standard baseline, traditional tensor regression, and regular dropout, and assess the robustness of each method in the face of adversarial noise.

CIFAR-100 (Krizhevsky and G. Hinton 2009) consists of 60 000 32×32 RGB images in 100 classes, divided into 50 000 images for training and 10 000 for testing. We processed the data by centering and scaling the intensities of each image and then augmented the training images with random cropping and random horizontal flipping.

ResNet classification	**Top-1 accuracy**
FC	75.88 %
FC + dropout	75.84 %
Tucker	76.02 %
CP	75.77 %
Randomized Tucker	76.05 %
Randomized CP	**76.19** %

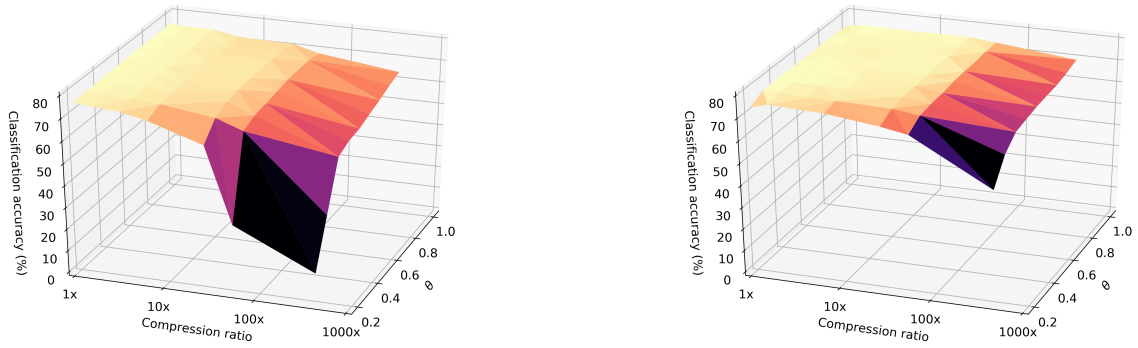
Table 5.2: Classification accuracy for CIFAR-100 with a ResNet and various regression layers for classification.

I compare the randomized tensor regression layer to full-rank tensor regression, average pooling and a fully-connected layer in an 18-layer residual network (ResNet) (He et al. 2016). For all networks, we used a batch size of 128 and trained for 400 epochs, and minimized the cross-entropy loss using stochastic gradient descent (SGD). The initial learning rate was set to 0.01 and lowered by a factor of 10 at epochs 150, 250 and 350. We used a weight decay (L_2 penalty) of 10^{-4} and a momentum of 0.9.

Classification results: Table 5.2 presents results obtained on the CIFAR-100 dataset, on which this method matches or outperforms other methods, including the same architectures without R-TRL. Tensor dropout method makes the network more robust by reducing over-fitting, thus allowing for superior performance on the testing set.

A natural question is whether the model is sensitive to the choice of rank and θ (or drop rate when sampling with repetition). To assess this, we show the performance as a function of both rank and θ in figure 5.3a. The reduction in rank is presented as the compression ratio $= \frac{\text{size of full tensor}}{\text{size of factorized cores and factors}}$. As can be observed, there is a large surface for which performance remains the same while decreasing both parameters (note the logarithmic scale for the rank). This means that, in practice, choosing good values for these is not an issue.

Performance as a function of rank and θ in replacement R-TRL: To illustrate the generality of the approach, which does not depend on the Bernoulli sampling, we perform a similar experiment with a different randomization: instead of using a Bernoulli random variable, we sample components with replacement according to a uniform sampling matrix (figure 5.3b). As for the Bernoulli case, there is a large surface for which performance remains the same while decreasing both parameters.



(a) Bernoulli R-TRL

(b) R-TRL with replacement

Figure 5.3: CIFAR-100 test accuracy, as a function of the compression ratio (logarithmic scale) and drop rate θ. There is a large region for which reducing both the rank and θ does not hurt performance.

Comparison with regular dropout: One question is whether the proposed tensor dropout induces more robustness than traditional dropout applied directly to the weights. To test this, we apply FGSM adversarial perturbations to each method, with varying magnitudes $\lambda \times 10^{-3}, \lambda \in \{1, 2, 4, 8, 16, 32, 64, 128\}$. We sample 1 000 images from the test set (Brendel, Rauber, and Bethge 2018). The models were trained *without* any adversarial training, on the training set, and adversarial noise was added to the test samples using the Fast Gradient Sign method. The results of which can be seen in figure 5.2. Our model is much more robust to adversarial attacks. Intuitively the method is able to leverage redundancies in the latent subspace, without creating holes in the weights, unlike dropout. In addition, since the randomization is used during training, this forces the latent decomposition to be over-complete and account for noise, thus rendering the model more robust to perturbation.

5.2.7 Conclusion

We introduced tensor dropout, a novel randomized tensor decomposition, suitable for end-to-end training of tensor regression layers. Adding stochasticity on the rank during training renders the network significantly more robust and leads to better performance. This results in networks that are more resilient to noise, both adversarial and random, without any addition such as adversarial training. Our results demonstrate superior performance on a variety of real-life, large-scale challenging tasks, including MRI data and images, as well as increased robustness.

5.3 High-order Kronecker machines for multi-modal learning

5.3.1 Introduction

In many medical settings we can obtain observations from different sources. Each source can contain unique information about the property under study. Our objective is to combine these sources to leverage all available information. This is vital for learning an accurate representation, since no single metric sufficiently describes an individual's health status. However, combining these different sources is challenging and most medical analyses are limited to the use of only a single source or dimension. Some advanced analyses model each source separately and then linearly combine (i.e. superposition) the individual outputs to create a final output. A major limitation is that this does not capture nonlinear interactions between features. A linearly-restricted model of a biological system can overlook important underlying processes.

Interaction mechanisms in biological systems are often non-linear and multivariate (Coffey 1998; Shafer 1995). Combining multiple modalities together can help reconstruct a latent representation of the health configuration of the individual. Learning labels from that representation is more powerful, as richer representations will lead to more accurate predictions of diseases and conditions. The data is already available yet underutilised, partly due to a lack of suitable analysis methods.

Multi-modal work can be classified based on where the mode fusion takes place. One strategy is to concatenate the raw data from all modalities and learn a representation from that merged data (Havaei et al. 2017; Pérez-Rosas, Mihalcea, and Morency 2013). In this case, all the learning takes place after the modalities are combined. At the other extreme, all the learning can be done prior to fusion, whereby the final output is calculated based on a voting or averaging scheme from multiple models that each operate on the different modalities (N. Liu et al. 2013; Nojavanasghari et al. 2016). The third and most diverse class of methods are those that first have modality-dependent learning, then fuse the learned representations together and finally learn a joint output (Veličković et al. 2016; Y. Peng et al. 2017).

This can be further divided into observations of a property using the same modality, for the same scene but captured from different positions (multi-view learning) and observations of a property using two different modalities (multi-modal learning).

Multi-modal learning often requires more general representation than multi-view learning, as the input structures can be of varying forms. Multi-modal methods are inherently multi-view, but the reverse does not necessarily hold.

However, most of these models are linear combinations of multimodal features and do not capture higher-order interactions. A major challenge in modelling higher order interactions is exponential growth in feature dimensions. Factorising the weight tensors is a useful compromise that can significantly reduce the computational complexity by restricting the space to a low-rank manifold while maintaining the expressive freedom for higher-order interactions.

Factorization machines (Rendle 2010) were not originally presented as a multi-modal learning tool, but its general nature could easily be built upon to account for multimodal data. That is precisely what has been done in Exponential machines (Novikov, Trofimov, and Ivan V. Oseledets 2017) and Multi-view FMs (Cao et al. 2016). In the context of brain imaging, Anderson et al. (2014) learn a factorised latent brain model using multimodal input features.

Here I seek inspiration from and build on these recent methods to develop a method for modelling high-order interactions between multiple modalities. I model the interactions using a factorised Kronecker product to reduce the number of parameters. This constrains the learning space which leads to improved training. Specifically, the contributions are:

- I present high-order Kronecker machines, a method to learn a latent representation of the individual from multi-modal brain images.
- I model a low-rank representation of interactions between learned features of each modality. This reduces the number of parameters, which has been shown to improve generalisability (5.2).
- To demonstrate that the whole framework can be learned end-to-end in a multi-modal settings in a large medical imaging application to outperform single-view and simple multi-view methods.

5.3.2 Kronecker machine

In this section, I will describe the Kronecker machine to model higher order interactions between multiple modalities, using a factorised Kronecker product. Factorisation allows us to learn representations in a low-rank tensor space, significantly reducing the computational cost to make the models tractable.

The Kronecker machine takes as input a set of tensors $(\mathcal{X}_1, \mathcal{X}_2 \dots \mathcal{X}_N)$, that are the learned representations of different modes, such as T1-weighted or T2-weighted MRI, fMRI, genetic data etc, as shown in figure 5.4. The model is described as the Kronecker product of all the tensors regressed with a weight tensor $\mathcal{W}$:

$$\hat{\mathbf{y}} = \mathcal{W} \circ (\mathcal{X}_1 \otimes \mathcal{X}_2 \otimes \ldots \otimes \mathcal{X}_N) \tag{5.11}$$

where $\mathcal{W} \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_N \times l}$ are the learned weights and l is the dimensionality of the output. Although the model can, in theory, learn from raw data, a representation of the data, modelled using a method appropriate for that particular might work better in practice.

The Kronecker product of raw data or sufficiently rich representations of individual inputs will be intractably large, in most real-world cases. $\mathcal{O}(d^n)$ for n modalities, each of dimension d. To alleviate this, we can initialise $\mathcal{W}$ as factorised representation where $\mathcal{W} = \mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times \cdots \times_N \mathbf{U}^{(N)}$

With this factorisation the model can be approximated as:

$$\hat{\mathbf{y}} = (\mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times \cdots \times_N \mathbf{U}^{(N)}) \circ (\mathcal{X}_1 \otimes \mathcal{X}_2 \otimes \ldots \otimes \mathcal{X}_N) \tag{5.12}$$

which can be re-written as:

$$\hat{\mathbf{y}} = \mathcal{G} \circ [(\mathbf{U}^{(1)} \times \mathcal{X}_1) \otimes (\mathbf{U}^{(2)} \times \mathcal{X}_2) \otimes \cdots \otimes (\mathbf{U}^{(N)} \times \mathcal{X}_N)] \tag{5.13}$$

This captures the N^{th}-order interactions between the inputs but does not explicitly model $< N^{th}$-order interactions. This is clear when inspecting the $N = 2$ case, where all weights correspond to products of both modalities - and no weights applied to only one or none of the modalities. This can be fixed with a simple trick, if we add an additional element to the low-rank representation of each factor with a constant value of 1, we can capture the 0^{th} and 1^{st} order interactions. The resulting tensor is then defined as: $\mathcal{D}_n = (1, \mathbf{U}^{(n)} \times_n \mathcal{X}_n)$

And then the final model with factorised weights and $< N^{th}$-interactions is written as:

$$\hat{\mathbf{y}} = \mathcal{G} \circ (\mathcal{D}_1 \otimes \mathcal{D}_2 \otimes \cdots \otimes \mathcal{D}_N) \tag{5.14}$$

The factorisation reduced the complexity to $\mathcal{O}((k+1)^n)$ for rank k decomposition,

which is significantly lower than composing the full tensor with rank d.

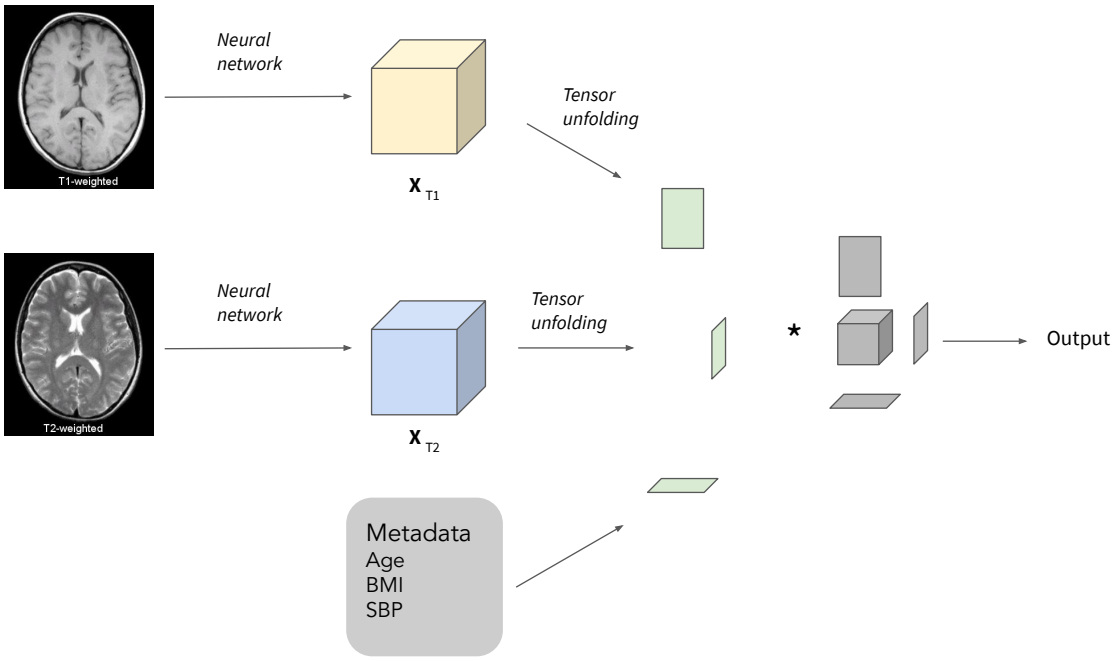


Figure 5.4: The Kronecker machine learns a joint factorised model of the interactions between multiple sources. Here three sources are depicted; e.g. T1-weighted and T2-weighted MRI scans and metadata. Note that in the experiments done here only the image inputs are used. The images are fed into separate 3D-CNNs that learn individual representations. The interactions between these representations are learned by the high-order Kronecker machine. A low-rank constraint through tensor factorisation reduces the size of the high-order interaction space making the operation far more efficient.

5.3.3 Experiments

I test the model in two phases. First, I use it to learn a toy image classification task (MNIST) and compare it with baseline methods to validate that multi-modal learning is taking place. This is followed by a test on multi-modal UK Biobank imaging data to investigate its performance in real world medical settings.

MNIST benchmark. To test the effectiveness of Kronecker machine integration of modalities I split the images into left and right as described previously (Andrew et al. 2013). This artificial multimodal task serves as a positive control. Models that are able to effectively use both halves of the images will perform similarly to a model that has access to the full image. Defining the three baselines is therefore straightforward: two separate models were trained for either side, and a third model was trained on the entire 28×28 image. A Kronecker machine is then set up in

the hypothetical scenario where the spatial relationship between the left and right image halves is unknown. The objective of the model is to learn the relationships between the two sides without any prior knowledge about the spatial relationship between the two halves.

The baseline models are simple CNNs with two convolutional layers, ReLU activation and max pooling, as described in the Pytorch MNIST demo (Paszke, Gross, Massa, et al. 2019). The left- and right-only configurations follow the same general configuration and differ only in the sizes of the activation tensors (and therefore, equivalent weight tensors). The final set of activations is half the size of the full CNN.

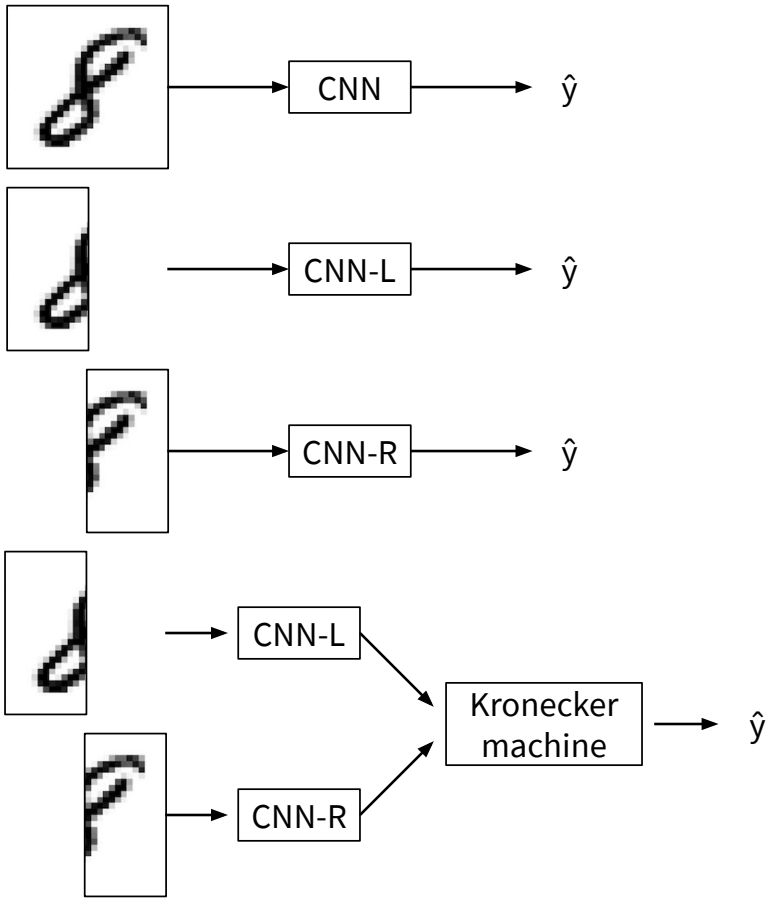


Figure 5.5: Overview of models used for comparison. The full-image CNN (top row) had access to the entire image while the CNN-L and CNN-R (rows two and three) had access to only the left and right sides of the image, respectively. The final model combines the output from CNN-L and CNN-R in a higher order Kronecker machine. Note that the Kronecker machine is trained end-to-end.

For the Kronecker machine, the final softmax activations were removed and the

layer instead fed into the fusion model as shown in figure 5.5.

All models were trained from random initialisation and optimised using Adam (Kingma and Ba 2015) with an initial learning rate of 0.01 for the Kronecker machine and 1 for the other models. All used a learning rate scheduler as described in the Pytorch MNIST demo (Paszke, Gross, Massa, et al. 2019) and were trained for 15 epochs.

Age prediction from T1 and T2 weighted brain MR images in UK Biobank. The next set of experiments explore the Kronecker machine's ability to combine information from two medical imaging sources: T1-weighted and T2-weighted MRI scans. It tests whether the model is able to leverage the interactions between the two sets of scans. These experiments are done under the assumption that the information between the two modalities includes interactions that are both linear and nonlinear.

I use the same T1-weighted brain MRI data as discussed in section 3.2. The T2-weighted scans were available for the same individuals and followed identical preprocessing as for the T1-weighted images. The resulting data gave inputs of 19 379 samples, each with two modalities, T1-weighted scans $182 \times 218 \times 182$ and T2-weighted scans of dimension $182 \times 218 \times 182$.

The initial networks for the T1-weighted and T2-weighted are identical and follow the 3D-Resnet architecture described in chapter 3. However, the activations from the final convolutional layer were not regressed into an output but instead input into the Kronecker machine, that factorised the Kronecker product using Tucker decomposition with a rank of 64 for all dimension. For implementation, the Khatri-rao product is used, with the batch dimension kept independent and not factorised. These final activations are of size $6 \times 7 \times 6$ for each of the two modalities.

For the single channel models, the activation tensors were flattened and a fully-connected layer convert them to a final output. The concatenated baseline followed identical architecture except the two modalities are concatenated at the input along a new dimension, resulting in a $2 \times 182 \times 218 \times 182$ input tensor. This is analogous to the standard representation of RGB colour images. For all experiments I split the samples at random into train, validation and test sets. The three sets contained 11 520, 3 847 and 3 847 samples, respectively.

Biological sex prediction from T1 and T2 weighted brain MR images in IXI. A second set of MR experiments was performed on the IXI brain image dataset (*IXI Dataset* n.d.). This set of 582 individuals with both T1 and T2-weighted brain MRI was used to test the model on a classification problem (biological sex) in a different population than that seen in UK Biobank experiments. The model was trained from random initialisation. The data was pre-processed using the methods described in Chapter 3, briefly: padding and/or cropping images to size $182 \times 218 \times 182$, per-image histogram intensity normalisation to zero mean and unit variance. T1 and T2-weighted scans from 480 individuals were used for training the model and scans from 102 individuals were used for testing.

Implementation details: Single channel 3D-Resnet models were trained as described in Chapter 3. Kronecker machine models were trained for 30 epochs using Adam (Kingma and Ba 2015) with a mini-batch size of 8 with an initial learning rate of 10^{-3} and weight decay of 5×10^{-5}. The learning rate was reduced by a factor of 10 at epoch 25.

5.3.4 Results

MNIST

On the handwritten digit (MNIST) toy problem (table 5.3), the CNN with access to the original full-image performs best with an accuracy of 99.31 %. Restricting the view to either only the left or right side of the images reduces the accuracy by around 4 percentage points. This is likely due to the model failing when one side of a written digit is ambiguous. For example, the left hand sides of the digits 4 and 9 can be identical in certain writing styles. Similarly, the right hand sides of the digits 3 and 8 can be attributable to either digit depending on writing style. The position is equally important as a shift to the left or right can make the problem of identification trivial, e.g. a digit 3 shifted sufficiently to the right will not be confused with a digit 8.

The Kronecker machine, combining the left and right sides sources of the images, performed significantly better than either side of the images alone. However, it did not perform as well as the full-image CNN set-up. Although both models had access to all pixels, the experiments were not identical. The original image contains implicit structural information since the two halves are connected in the relevant spatial dimension, which is lost when the images are bisected. Results on this control experiment suggest that the Kronecker machine can leverage the two data

sources.

Table 5.3: MNIST classification accuracy.

Model	Top-1 accuracy
CNN - full image	**99.31 %**
CNN - left only	95.20 %
CNN - right only	95.46 %
Kronecker machine	98.19 %

UK Biobank

The Kronecker machine predicted age with a MAE of 2.39 years on the test set (N =3 847) of UK Biobank individuals (table 5.4). This is a lower error than the 3D-ResNet trained on either T1-weighed or T2-weighted MRI scans individually (lowest MAE of 2.68 on the T1-weighted scans). The Kronecker machine also outperformed the 3D-ResNet that combines T1 and T2-weighted scans via concatenation in the input.

Table 5.4: Age prediction accuracy from brain MRI (UK Biobank). Reported here is the MAE of predictions on the test set.

Model	T1-weighted	T2-weighted
3D-ResNet	2.68 years	2.74 years
3D-ResNet - concatenated	2.45 years	
Kronecker machine	**2.39** years	

In classification experiments on IXI brain MRI scans, the Kronecker machine outperforms the other models (last row in table 5.5). Here, the Kronecker machine took inputs from two 10-layer 3D-ResNets and predicted with an accuracy of 92.31% on the test set. The 3D-Resnet-10, which is directly comparable, was only able to predict with accuracies of 89.41% and 90.38% with individual T1 and T2-weighted scans, respectively. The deeper 18-layer 3D-Resnet performed better on T1-weighted scans, but did not match the performance of the Kronecker machine.

Table 5.5: Biological sex classification accuracy from brain MRI (IXI test set, N=480).

Model	T1-weighted	T2-weighted
3D-ResNet-10	89.41 %	90.38 %
3D-ResNet-18	91.35 %	84.54 %
Kronecker Machine	**92.31 %**	

5.3.5 Discussion

The Kronecker machine shows improvements over the other methods. MNIST experiments show that it is combining information from the two sources (halves) of the image for improved performance. UK Biobank tests show that the Kronecker machine performs better than concatenating the two channels at the input, similar to RGB images. The final IXI results further demonstrate benefits of the model. The reported accuracy is lower on IXI data than UK Biobank due do a significantly smaller training set.

Toy data experiments on MNIST suggest that the Kronecker machine leverages both sources by outperforming both CNNs that used only a single side of the image. The choice of factorisation rank was not explored as part of this study. As we saw in Section 5.2, rank selection can have a significant impact on the network's ability to model the data. If the tensor factors have a rank that is insufficient, the model may be too restricted to capture the necessary relationships. Conversely, larger ranks increase computational complexity to consume the operational and memory budget. Here, I have heuristically selected a rank of 64 and used throughout the study. The optimal hyperparameter is likely dependent on the dataset and the intrinsic dimension of the objective (C. Li et al. 2018). Optimising the rank conditioned on the data will likely result in improved performance.

Related to the previous argument is the Kronecker machine input selection. In these experiments I have used the state-of-the-art computer-vision models[1] (a two-layer CNN for MNIST and a 3D-Resnet for brain imaging tasks). The ability of the Kronecker machine, and any mid- or late-stage fusion method, is highly dependent on the quality of the learned representations returned from the CNNs. Training the entire model end-to-end does, theoretically, allow the source-specific models to optimise for representations that are best suited for the fusion model. Nevertheless, all models are limited in their intrinsic modelling capacity.

A limitation of this study is lack of comparison with more multi-modal integration models. Without direct comparisons, the relative benefits of the Kronecker machine over other fusion models cannot be stated. Combining multiple MRI modalities together is known to provide better results over single-modality methods for medically related tasks, including brain tumour segmentation (Soltaninejad et al. 2018), identifying neurological biomarkers for schizophrenia (Sui et al. 2018) and characterising

[1]Technically, CNNs with more than two layers is optimal for MNIST, here I refer to CNNs as the class of methods that represent the state of the art.

peripheral inflammation (Schrepf et al. 2018). However, these studies do not model the nonlinear interactions between the source modalities as done here. For future work, a controlled comparison between Kronecker machines and nonlinear fusion methods, e.g. Exponential machines (Novikov, Trofimov, and Ivan V. Oseledets 2017) and Multi-view FMs (Cao et al. 2016), on an appropriate medical task will bring the Kronecker machine performance into context. However, the results shown here demonstrate the feasibility of the Kronecker machine factorisation technique to combine multi-source medical data.

5.3.6 Conclusion

Combining multiple sources of information remains a challenge due to the exponential number of interactions that have to be modelled. Kronecker machines provide a clever way of jointly modelling representations learnt by model-specific architectures to improve prediction performance.

Low-rank constraints provide another benefit in the fact that the model contains fewer parameters than the full-rank equivalent. This is particularly important when deploying models in practice. Real-world products have limited memory capacity that needs to be managed and shared with other applications. This is particularly relevant to medical applications as privacy limitations often prohibit patient data to be sent to a centralised server which can store large models. Smaller models are more likely to fit on edge or mobile devices that can run model inference on site, where the data is collected.

The results showcase the flexibility of the Kronecker machines. Their properties makes them agnostic to the choice of model used for processing source-specific inputs. They are also independent of the type of data being modelled, opening up the possibility of incorporating different medical data, such as wearable device recordings, genetic or metabolomic data to improve predictions. These practical factors make them versatile and worthy of further research.

Chapter 6

Deep learning for polygenic predictions

Risk stratification from genome sequences is a growing area of research with large potential implications for disease prediction and prevention. However, most current polygenic prediction methods lack the capability to model non-linear effects between genetic variants. It is believed that the majority of phenotypes are controlled by small compounding effects and interactions from genetic variants across the entire genome. Here, I present a novel approach that can predict traits from large genome sequences by leveraging both local and global interactions between genetic variants. The model is a deep network of growing locally-connected receptive field layers that capture an increasing proportion of the genome as the depth increases. In experiments, the model matched state-of-the-art performance on two large datasets: 1000 Genomes and UK Biobank.

6.1 Introduction

Estimating phenotypes from an individual's genotype is of high value to disease prediction and, ultimately, prevention. Although many traits show significant associations with Single Nucleotide Polymorphisms (SNP)s (Locke et al. 2015; Kunkle et al. 2019), recent findings suggest that a significant number of traits are a function of small effects from a large number of genetic variants that are distributed across the entire genome (Dudbridge 2013). Further studies have shown that modelling multiple variants that are by themselves not genome-wide significant can lead to improved prediction (Mavaddat et al. 2019). Therefore, the accuracy of map-

pings from the genome to phenotype is theoretically increased with the number of modelled genetic variants. This forms the first design motivation for the method introduced here.

The second factor considered is the inherent structure of the data. As the order of characters, words and sentences conveys meaning in text, the organisation of the genome implicitly contains information. As with text, this includes both local and global structure, from the arrangement of DNA bases in codons to entire haplotype blocks (Gabriel et al. 2002). However, this auxiliary information is often excluded in phenotype prediction methods.

Terminology
P-value thresholding at P_{thres} means to remove all variants with GWAS P-values of association that are less than P_{thres}.
Shrinkage: Penalise regression coefficients, typically with either L_1 (lasso) or L_2 (ridge) regularisation.
Linkage Disequilibrium (LD): SNPs in a locus that are highly correlated with each other will be overestimated in the PGS. LD pruning SNPs can reduce negative effects of LD by removing highly correlated variants that bias the model. However, excessive LD pruning can be detrimental to accuracy if variants that are slightly correlated, but encode for unique information are removed.

A common approach for risk prediction from genomic data are GWASs, as applied in chapter 4. Briefly, these studies typically investigate the associations between individual SNPs and trait. GWASs involve multiple univariate association tests on every available variant and is hypothesis free, although priors can be built-in for certain cases (Walters, Cox, and Yaacob 2019; Wallace 2020). Associations with individual SNPs are useful for many purposes, but they are weak signals for most traits outside of a limited number of rare diseases (Dudbridge 2016). Efforts have been made to combine the effect of multiple SNPs to improve heritability predictions, demonstrated by the use of linear mixed modelling to predict human height (J. Yang et al. 2010). Recently introduced, Polygenic scores (PGSs), also known as a polygenic risk scores, are the current prediction method of choice. The prototypical PGS formulation is a linear weighted sum of contributions from the selected SNPs (Wray et al. 2014) defined as "a single value estimate of an individual's genetic liability to a phenotype, calculated as a sum of their genome-wide genotypes, weighted by corresponding genotype effect size estimates derived from GWAS summary statistic data." (Choi, T. S.-H. Mak, and O'Reilly 2020), or

more precisely:

$$\hat{y} = \sum_{j=1}^{P} x_j \hat{\beta}_j$$

where $\hat{y}$ is the risk score, j is the index of P SNPs. $\mathbf{x}$ is the vector of genetic markers for the individual and $\hat{\beta}_j$ is the weight assigned to SNP j, as estimated from the summary statistics.

More recent improvements to PGS, such as lassosum (T. S. H. Mak et al. 2017), have made use of lasso regularisation to outperform linear regression PGS estimations in both simulations and real-world risk stratification (J. Elliott et al. 2020). It is argued that penalised regression makes p-value thresholding redundant.

The aforementioned methods account for only linear interactions between genetic variants across the genome. Extensions beyond that are faced with three main hurdles. First, how to effectively use information from summary statistics about first order contributions. Second, to model nonlinear fusion between multiple genetic variants. Finally, if and how to select variants to include in the analysis. Here, I approach these challenges using a novel deep learning architecture for phenotype prediction from genomic data. Specifically, the contributions are as follows:

- Introduce receptive field networks: a deep neural network architecture that models nonlinear fusion between multiple variants and accounts for local and global structure in the genome.
- Demonstrate that receptive field networks match state-of-the-art performance in PGS prediction on multiple tasks in two large datasets: 1000 Genomes and UK Biobank.
- Present an intuitive way of incorporating summary statistics into deep learning models.
- Provide proof-of-concept analysis for interpreting the model to identify variants of high importance.

6.2 Receptive Field Networks

Here I introduce a novel deep learning architecture for PGS estimation: receptive field networks. The primary design criteria are to A) learn nonlinear interactions between variants, B) make use of the inherent structure of genome and C) leverage summary statistics from previous GWASs. The model seeks inspiration from convo-

lutional layers in computer vision, and the Wavenet (Oord et al. 2016) and ResNet (He et al. 2016) architectures. The design is motivated by the idea that initial layers will both leverage summary statistics and capture local interactions between variants that are located in the same region on the genome. As the network deepens, further layers will capture more global relationships between the lower-level features, eventually covering the entire input space to generate an output. The entire network is trained end-to-end using backpropagation.

The novelty of the method comes from combining residual priors in the input layer with local layers that build depth and learn interactions between variants. The residual priors resemble traditional residual layers for deep networks, but learn the deviation from the available GWAS summary statistics. Local layers connect proximate variants and representations together and learn a representation for that grouping.

6.2.1 Residuals priors

To incorporate summary statistics I modify the first layer of the network to learn deviations (residuals) from the summary statistics. Most neural network layers can be described in the framework of:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \mathbf{W}) + \mathbf{B}\mathbf{x} + b$$

in traditional networks $\mathbf{B} = 0$ and the layer output $\mathbf{y}$ is a function of only the layer input and the optional bias term b. In ResNets, $\mathbf{B} = 1$ and $\mathcal{F}(\mathbf{x}, \mathbf{W})$ denotes learned deviations from the identify function. This allows residual networks to be much deeper than traditional ones, as any redundant layers will tend to $\mathcal{F}(\mathbf{x}, \mathbf{W}) = 0$, allowing the information to pass through. I build on this idea in layers that I call *residual priors*. In this case $\mathbf{B} = \hat{\boldsymbol{\beta}}$, i.e. the corresponding genotype effect size estimate derived from GWAS summary statistics data. Therefore, the unmodified output (in the case when $\mathcal{F}(\mathbf{x}) = 0$) of the layer will be the weighted sum of the summary statistics and correspond to a classic linear PGS for that region. When summary statistics are not available, the layer can revert to a typical residual network mode of operation where $\mathbf{B} = 1$.

6.2.2 Local layers

The local layer connects regions proximate in the previous layer together. A schematic overview of this is shown in figure 6.1. There are many frameworks in which this layer can be described. It can be described of as a standard neural network layer where only nodes that are within a certain region are used as input. It can also be thought of as a 1D-convolutional layer where the weights are unshared. Each patch of input will have a unique set of learned weights. Finally, this is also a temporally-static interpretation of Wavenets (Oord et al. 2016). Wavenets are a generative model of raw audio that uses exponentially growing receptive fields to process audio signals at various timescales. The local layer is defined for a set of units, $\mathbf{x}_{subset}$, in the preceding layer as:

$$y_{subset} = \mathcal{F}\Big(\sum_i \mathbf{W}_{subset,i}\mathbf{W}_{subset,i} + b_{subset}\Big)$$

where $\mathbf{W}_{subset}$) and b_{subset} are the learned set of weights and bias unique for this subset and y_{subset} is a single unit in the layer.

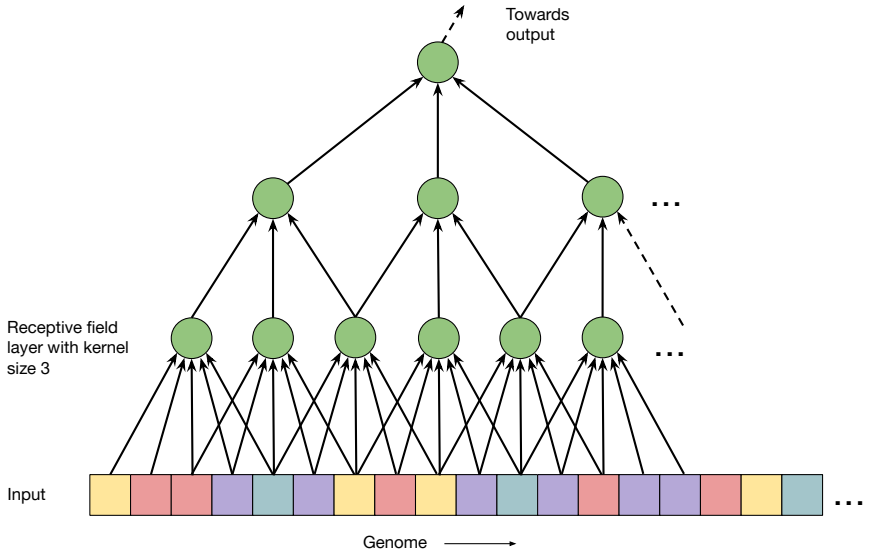


Figure 6.1: Schematic overview of a Receptive field network. Multicoloured boxes represent different variants in the input genome sequence. These are connected to receptive field layers. The size of the receptive field (kernel) in the first layer is 5 with a stride of two. The second layer has a receptive field of 3 and stride of two. Each subsequent layer connects a greater proportion of the input together and all parts are eventually connected in the output layer.

6.3 Data and set up

Two genetic datasets will be used to test and compare the models: The 1000 Genomes Project (Consortium et al. 2012) and UK Biobank (Sudlow et al. 2015a). Although they contain data from completely different populations and the variants recorded do not overlap completely, the data is organised in a standard way. In both cases, the genomic data is represented in a matrix where the rows correspond to the samples (individuals) and the columns to the variants. $\mathbf{X} \in \mathbb{R}^{N \times P}$ where N is the number of samples and P is the number of variants. In both datasets the hard-coded or "best guess" variant values are used. Therefore, each variant can take value of $0, 1, 2$, representing the number copies. For labels on geographic populations, I use data as reported in the 1000G and UK Biobank databases, with the UK Biobank recording self-reported data.

6.3.1 1000 Genomes dataset

The 1000 Genomes Project (Consortium et al. 2012) is a collection of low-coverage whole-genome sequencing of over 1 000 individuals from 26 geographical populations. The resources' objective was to capture the majority of SNPs at a frequency of 1%.

Here, I use this dataset for development and comparison. Therefore I follow the preprocessing described previously (Romero et al. 2017). This included: selecting variants with frequencies $\geq 5\%$ and with a linkage equilibrium $r^2 < 0.5$, and excluding sex chromosomes, giving a total of 315 345 variants in the processed data.

The task is to predict the ethnicity (geographic group) of an individual based on the genotype. By following the same task set up as done by Romero et al. (2017), I can compare the methods directly.

6.3.2 UK Biobank dataset

UK Biobank is a very large population-based cohort study of approximately 500 000 participants from the UK, aged 40-69 years old. At the initial data collection, a wide range of records were recorded including blood samples for biochemical tests, whole genome sequencing, various physical measurements and self-reported information, with their data linked to Hospital Episode Statistics. For a more detailed description of the data collection and processing we refer the reader to Sudlow et al. (2015a).

The UK Biobank genetic dataset with Coronary Artery Disease (CAD) labels contains 31 894 individuals where each sample $\mathbf{x} \in \mathbb{Z}^{487\,293}$ can take value of $0, 1, 2$, representing the number copies of each effective variant. I used UK Biobank's genotyped data and preprocessed by filtering variants with an INFO score > 0.999. This resulted in a total of 473 876 variants. To adapt and calibrate the model architecture initially to UK Biobank data, I performed a heuristic hyperparameter search on ethnicity prediction and used the best performing setup as a starting point for CAD hyperparameter tuning.

6.3.3 Benchmark models

To measure the performance of receptive field networks, I compare it against three baselines: linear regression on all tasks, Diet Networks (Romero et al. 2017) on ethnicity prediction in the 1000G dataset and a lassosum implementation (J. Elliott et al. 2020) for CAD prediction in UK Biobank.

The linear regression model provides a lower-bound for performance while remaining interpretable. The learned weights of the model are regularised with an L_2 penalty (also known as ridge regression), and a relatively high dropout ($p = 0.5$) to reduce overfitting on the training set. In cases where summary statistics are available, the weights of the linear model can be fixed to the corresponding variant effect size. In other cases, the weights can be learned from a hold-out training set.

Diet Networks (Romero et al. 2017) are a recently introduced deep learning approach for prediction from genomic data. Faced with the problem of an exponentially growing number of parameters as more variants are included, the authors fix the weights in the first layer of the fully-connected neural network to pre-calculated embeddings, similar to summary statistics. The rest of the network is trained end-to-end with both a supervised loss and a separate reconstruction loss that serves as an unsupervised representation learning signal. This method is a direct comparison for the ethnicity prediction task on the 1000G dataset. Finally, I compare the performance of the receptive field networks on CAD prediction with a state-of-the-art lassosum study (J. Elliott et al. 2020).

6.3.4 Set up and hyperparameters

For experiments on the 1000G data, I use a four-layer network. The first three layers are local layers with receptive field (kernel) sizes 37, 37, 4 and strides 3, 3, 1 in the first, second and third layer, respectively. The final layer is fully-connected and

activated with a softmax to generate the logits used for prediction. The networks were optimised by minimising the cross-entropy loss between the predicted and true distributions using Adam (Kingma and Ba 2015) for 300 epochs and mini-batch size of 128 with an initial learning rate of 10^{-4}, decaying by a factor of 10 at epochs 30, 50, 100 and 200. To reduce overfitting and improve generalisation to unseen data I applied an L_2 weight decay of 10^{-3} on all parameters during training. For all experiments I split the samples at random into train and test sets that contained 2 760 and 690 individuals, respectively. Summary statistics were not used for the ethnicity predictions.

Heuristic hyperparameter search experiments on UK Biobank data resulted in a slightly different architecture. The first three layers are local layers, all with a receptive field (kernel) size of 64 and stride of 32. The final layer is fully-connected and gives a scalar output representing the CAD risk prediction. The networks were optimised by minimising the cross-entropy loss between the predicted and true distributions using Adam (Kingma and Ba 2015) for 300 epochs and mini-batch size of 128 with an initial learning rate of 10^{-2}, decaying by a factor of 10 at epochs 30, 50 and 200. An L_2 weight decay of 10^{-3} was applied on all parameters during training. For all experiments I split the samples at random into train and test sets that contained 23 920 and 7 937 individuals, respectively.

Summary statistics for CAD predictions were sourced from CARDIoGRAMplusC4D (Nikpay et al. 2015). That data on coronary artery disease / myocardial infarction have been contributed by CARDIoGRAMplusC4D investigators and have been downloaded from www.cardiogramplusc4d.org.

6.4 Results

6.4.1 Task visualisation

Using a similar approach as in chapter 3, I visualise the intrinsic structure of UK Biobank genetic data using unsupervised manifold learning. UMAP (McInnes and Healy 2018) learns the embedding shown in figure 6.2. We can see clear structure in the embedding of ethnicity when individual samples are labelled. Figure 6.3 further highlights the clear separation of classes, with geographically proximate and overlapping groups, such as Pakistan and Bangladesh, mapping so proximate regions in the genetic embedding.

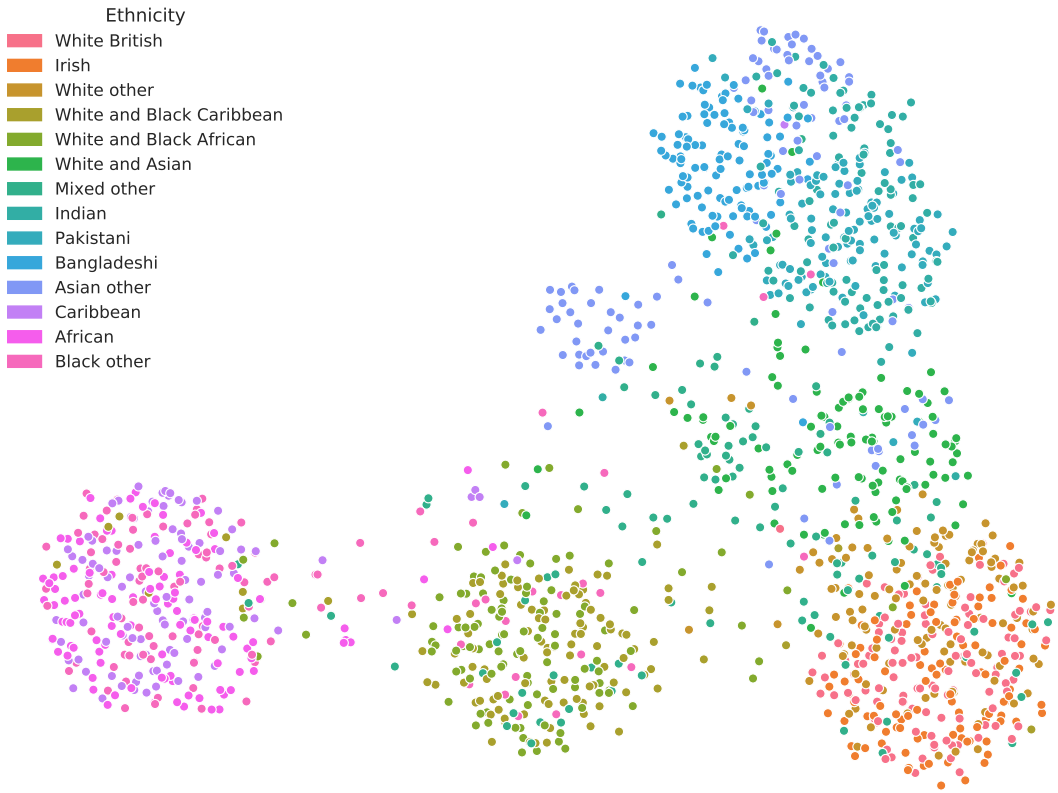


Figure 6.2: UMAP of UK Biobank data using self-reported ethnicity coding. Each point represents an individual, the labels were not available to UMAP during training.

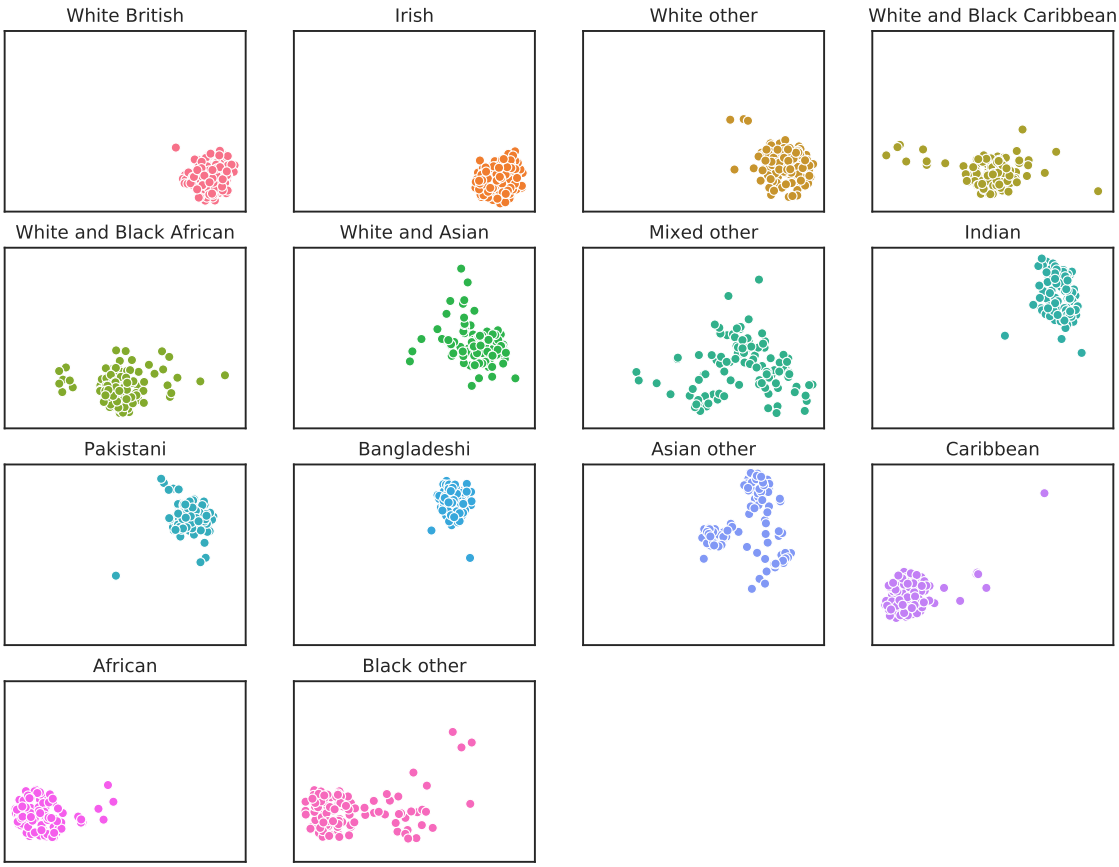


Figure 6.3: UMAP of UK Biobank data using self-reported ethnicity coding.

6.4.2 Experiments on 1000G data

The linear model performs very well, as shown in table 6.1, achieving a top-1 error of 4.06% (±0.94%) (mean ± standard deviation of three runs). Diet Networks perform worse with an error rate of 7.44%(±0.45%), here mean and standard deviation are from five runs, as reported in the original study. Top-3 error rate was not reported. The performance of Receptive Field Networks is between the linear model and Diet Networks in Top-1 accuracy.

Since the models output the logits of all classes it is possible to rank the Top-3 error rate, i.e. the number of samples in the test set where the correct class was not one of the three highest logit outputs. Receptive Field Networks outperformed the linear model in this setting.

Table 6.1: Comparison of models on the ethnicity prediction task in 1000 Genomes database. The top-1 and top-3 error rates are reported. Top-k error is the proportion of samples in the test set where the correct label was not in the top-k highest scoring outputs of the model. The top-3 error rate is not reported in the Diet Networks study.

Model	Top-1 Error (%)	Top-3 Error (%)
Linear	$\mathbf{4.06 \pm 0.94}$	0.62 ± 0.16
Diet Networks (Romero et al. 2017)	7.44 ± 0.45	N/A
Receptive Field Networks	5.52 ± 0.32	$\mathbf{0.47 \pm 0.20}$

6.4.3 Predicting CAD in UK Biobank

Before predicting CAD, I test the model's ability to learn from UK Biobank data on the same task as for 1000G: ethnicity prediction. The Receptive field networks outperformed the linear model (table 6.2), with a top-1 error rate of 6.95 % and 7.59 %, respectively. This demonstrates that the Receptive field networks are able to learn a simple task from UK Biobank genetic data. The results suggest that with more data (number of samples), leads to improved performance as the Receptive field networks outperform the linear model on the UK Biobank but not on 1000G.

For CAD risk stratification with a case-control dataset, the receptive field networks achieve an AUC of 0.59 (table 6.3). This is substantially lower than the state-of-the-art Lassosum method, which achieves an AUC of 0.63.

Table 6.2: Ethnicity prediction in UK Biobank. The top-1 error represents the proportion of samples in the test set (N=7 937) for which the model predicted an incorrect class, out of the 14 labels available.

Model	Top-1 Error (%)
Linear	7.59
Receptive Field Networks	**6.95**

Table 6.3: CAD prediction in UK Biobank. Reported here is the AUC of the receiver operating characteristic curve for predictions of samples in the test set (N=7 937).

Model	AUC
Lassosum (J. Elliott et al. 2020)	**0.63** (95% CI, $0.62 - 0.64$)
Receptive Field Networks	0.59 (95% CI, $0.57 - 0.61$)

6.4.4 Analysis of receptive field activations

I conducted an activation analysis of the receptive field networks to investigate the contribution of different regions on the final output. The objective of this analysis is to see whether the network is relying on a small number of localised variants or utilising effects from across the genome.

The results are presented in figure 6.4. The x-axis plots the distribution of the mean activations of the units of the first layer of the receptive field network, when predicting on the test set of 1000G data. Here, activation refers to the output of the unit after the activation function has been applied. As can be seen on the top bar plot, a significant number of regions have a mean of 0. In those cases, the ReLU action has cut-off the negative output of the unit and effectively stopping a signal being passed through these units. One way to interpret those is that the network has learnt to 'ignore' these zero-regions. The majority of the regions, however, do return a nonzero signal. A region returning a nonzero signal is not a sufficient condition for the region to be classed as informative. If the region returns a constant signal independent of the input, then it contains no information. The variation in the signal is visualised on the y-axis. There is a clear distribution of standard deviations, indicating that the output of the different regions varies with the input.

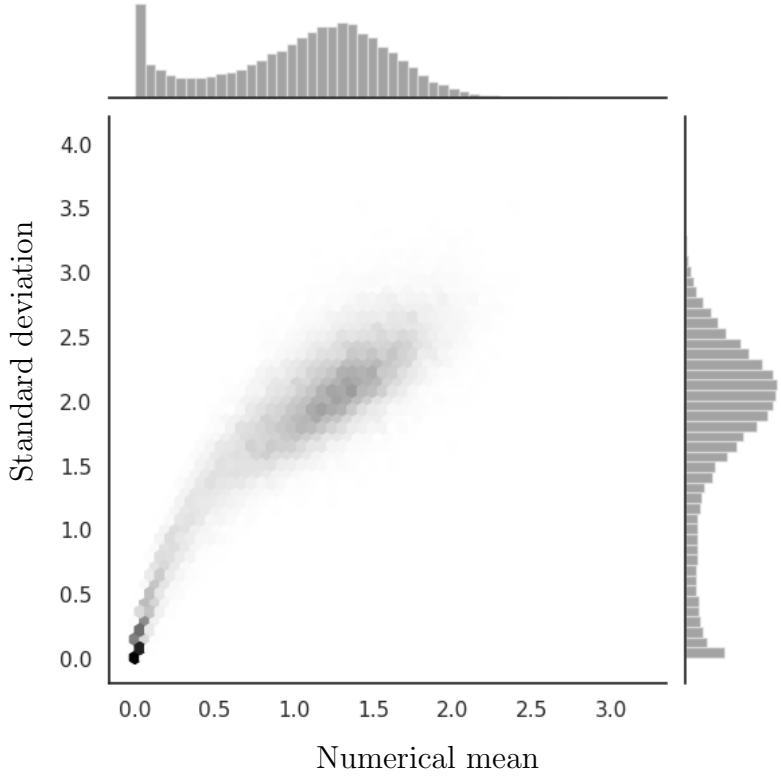


Figure 6.4: Activation analysis of the first layer of a receptive field network tested on 1000G data. The distributions plotted are for the mean (x-axis) and standard deviation (y-axis) of the outputs of the first layer (after ReLU activations). The wide distribution and relatively few zero values indicate that the network is varying with a wide range of regions in the genome.

6.5 Discussion

The Receptive field networks are a promising deep neural network architecture designed to leverage local and global structure of the genome and nonlinear interactions between variants. I showed that it outperforms both previous deep learning methods and, given enough data, linear methods.

Given more data, in the form of increased samples and variants, the Receptive field networks were able to outperform linear methods on the task of classification by geographic ethnicity. Deep neural networks, including receptive field networks, have significantly more capacity than their linear counterparts. Although the added capacity allows them to capture more complex interactions, it comes with the added cost of more difficult training.

Activation analysis revealed that the network was varying with a large number of variants across the genome. However, variation does not guarantee influence on prediction accuracy; a part of these dynamics is due to noise but the extent of which is unknown. Another limitation is that this analysis is done only on first layer, regions can be zeroed in deeper layers, therefore nullifying any contribution from those regions. More advanced activation analyses could potentially be used to interpret and identify variant contributions.

To what extent the model makes use of structure was not explicitly demonstrated. Further work is needed, possibly by permuting the genome to break structure as a negative control. Potentially, high LD pruning might be removing local structure. Kernel size and stride were heuristically selected with a small set of hyperparameter observations. A comprehensive hyperparameter search will inevitably lead to improvements in prediction accuracy. Selection for the two hyperparameters is not independent. When stride $> \frac{\text{Kernel size}}{2}$, inputs on the kernel periphery are used as inputs for two adjacent units in the following layer. This leads to an overrepresentation of variants on the kernel peripheries, which can lead to bias.

A limitation of the input definition used in this study is that variants are not evenly spaced relative to their absolute location on the genome, i.e. one receptive field might span a region covering M basepairs in the genome, while the next receptive field covers a different region spanning N basepairs. However, both regions will take as input the same number of genotyped variants.

Previous attempts at using deep learning for heritability or risk prediction from genetic data have focused on convolutional neural networks (Laksshman et al. 2017;

Abdollahi-Arpanahi, Gianola, and Peñagaricano 2020). This is unsurprising, as deep learning has revolutionised computer vision, with CNNs as the go-to modelling method. The resulting fame of CNNs and abundance of available code implementations make them the obvious choice for a first attempt at deep learning for genomics, but this can be misguided. The CNN architecture is not designed for genomic data. Convolutional kernels are effective for capturing identical patterns that are observed at different spatial locations in a data sample. For example, in image recognition early layers learn to detect localised patterns, such as edges, while deeper layers combine the outputs from the lower layers to identify objects and the final layers unify the object semantics into a final output (LeCun, Bengio, and G. Hinton 2015). The convolutional layers used in CNNs are translationally equivariant (Kondor and Trivedi 2018). This quality allows the models to recognise scenes independently of where the objects of interest appear - the location of a cat in an image does not change whether the image contains a cat. However, this is fundamentally different to genomic data, where the location of specific variants is constant across all data points. For images, inverting a pixel or other such small perturbations does usually not change the true interpretation and CNNs are relatively robust to such changes, as we see in chapter 5. Genomic data behaves differently: changes to a single variant (SNP) can have profound effects on the phenotype or, more commonly, changes to a small number of variants across the genome.

Genomic data is more similar to text than to images. Although all three are constructed with hierarchical semantics, vision is more pattern-based than the one-dimensional data types. Small changes to text can completely change the meaning: with only one removal and one character switch *she has one apple* can be changed to *she has no apple*. CNNs struggle with these semantics and have not established themselves as state-of-the-art models for natural language processing.

A class of deep learning architecture known as Transformers might be a solution. Transformers are extremely versatile. Originally proposed for natural language processing (Vaswani et al. 2017), they have been adapted with relatively few changes to work on other structurally different data, including images (Anonymous 2021).

Note that *ethnicity* is in general not well-defined and includes both genetic and environmental (social) factors and is arguably a flawed determinant for disease causation (Collins 2004). However, the concept, as reported in 1000G and UK Biobank data, includes a signal for ancestral geographic origins and genetic subgroups. Although this labelling is imperfect, these geographic origins can be clustered, as the

unsupervised manifold learning demonstrates. Learning these clusters serves as a positive control to indicate that these clusters can be learned in a supervised system, but the learned model serves no further purpose.

6.6 Conclusion

I have presented Receptive field networks, a deep learning approach that models local and global genome structure in addition to nonlinear fusion interactions between variants. Results show that the prediction advantage of Receptive field networks over linear models increases with available training data. This suggests that the network can pick up weaker signals in large population cohorts, demonstrating the potential for future applications as sequencing becomes more readily available - a trend that is likely to continue. A further comparison on CAD prediction with state-of-the-art methods showed that Receptive field networks return similar results. Accurately predicting the genotypic contribution to CAD risk is important for prioritising individuals who would benefit from risk reduction through treatment or intervention.

Building on this concept of localised connections with exponentially growing receptive fields could provide a platform for architectural variations. With greater computational resources and hyperparameters optimisation, it is likely that the performance can be improved even further than seen here. Moreover, adding attention mechanisms to help focus the relationships between layers is an interesting research direction. Taken together, these results demonstrate novel deep learning approaches show merit as a predictor for serious health conditions from large genome sequences.

Chapter 7

Conclusions

In this thesis we have seen how deep learning can be used to learn representations for health outcome prediction. First, a deep neural network that learns a biomarker from brain structural MRI that provides a useful measure for investigating systemic health and can augment neuroradiological research. We also saw how deep learning can predict polygenic risk scores with performance comparable to, and in some cases better than, that achieved in clinical settings. These advancements are made possible by my developments in deep neural networks and tensor methods. Tensor dropout improves generalisability and robustness to noise, and Kronecker machines can combine multiple imaging modalities for improved predictive capabilities. Together, these inventions and discoveries highlight the capabilities of deep learning for health outcome prediction, and set the scene for real-world applications that can significantly improve health and quality-of-life.

One application of these methods is polygenic risk scoring. More accurate risk scores allows genetic determinants for ill-health to be identified early and individuals to be stratified for proactive measures for risk factor remediation. Not only does this lead to improved personal health outcomes, it also saves valuable healthcare resources that otherwise would have been spent on reactive treatment.

Taking MRI into account can provide complementary benefits to polygenic risk scoring. Brain structure is not wholly genetically pre-determined but is influenced by lifetime exposure to environmental and lifestyle risk factors. Many diseases and conditions are caused by or associated with such exposures, some of which manifest as changes in brain structure. Combining genome sequences and brain MRI data in a multi-modal system can improve prediction capabilities and would be a natural

next step for the proposed Kronecker machine.

With further validation in new populations, these methods can serve in clinical environments as decision-support tools. Further additions, such as a reliable uncertainty measure on the predictions, can turn the systems such as the ones presented into a part of a more complete automated healthcare framework, where decisions are optimised based on previous patients. A reinforcement learning system with that level of autonomy carries both enormous potential for improvements to efficiency and big risks. A poorly calibrated model that learns nonsense actions can cause almost unlimited damage. Trust can be built over time with interpretability, similar to what I did with the MRI predictions. Interpretations can also guide future research by shedding light on unknown relationships and aid in hypothesis forming.

While this work is principally focused on supervised learning, recent developments in deep self-supervised learning, where models learn representations without labels, have shown great promise. This is particularly applicable to rare diseases, as patterns from related conditions can be leveraged to improve learning. Moreover, the disease classification scheme devised by humans does not perfectly reflect the inherent biological relationships between disorders. Self-supervision might help us learn more about the clusters of diseases, create finer ways of distinguishing between two seemingly related, but biologically unrelated, diseases. A more comprehensive overview of disease hierarchy and their relationships can aid identification of treatments that can be adapted to related conditions. It is relatively straightforward to use the models I presented in such self-supervised learning systems. The structure of the models can be kept unchanged and only the loss function changed.

Using multi-modal methods to combine different types of data offers an exciting future research avenue. Using high-order Kronecker machines (or similar methods), it is possible to leverage both imaging and genomic data. In theory, any relevant data can be combined in the prediction, including electronic health records, activity data from wearable devices and biochemical measurements. UK Biobank makes such analyses possible with its comprehensive data collection and detailed labels.

These contributions clearly demonstrate the benefits of using deep learning for health outcome prediction in both research and clinical settings. I hope my work inspires other scientists and enables practitioners to build applications with great and far-reaching impact.

References

Abdollahi-Arpanahi, Rostam, Daniel Gianola, and Francisco Peñagaricano (2020). "Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes". In: *Genetics Selection Evolution* 52.1, pp. 1–15.

Acharya, U Rajendra, Hamido Fujita, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Muhammad Adam (2017). "Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals". In: *Information Sciences* 415, pp. 190–198.

Akkus, Zeynettin, Alfiia Galimzianova, Assaf Hoogi, Daniel L. Rubin, and Bradley J. Erickson (2017). "Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions". In: *J. Digit. Imaging*.

Anderson, Ariana E., Pamela K. Douglas, Wesley T. Kerr, Virginia S. Haynes, Alan L. Yuille, Jianwen Xie, Ying Nian Wu, Jesse A. Brown, and Mark S. Cohen (2014). "Non-negative matrix factorization of multimodal MRI, fMRI and phenotypic data reveals differential changes in default mode subnetworks in ADHD". In: *NeuroImage*.

Andrew, Galen, Raman Arora, Jeff A. Bilmes, and Karen Livescu (2013). "Deep Canonical Correlation Analysis". In: *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*.

Angermueller, Christof, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle (2016). "Deep learning for computational biology". In: *Molecular systems biology* 12.7, p. 878.

Angevaren, Maaike, Geert Aufdemkampe, HJJ Verhaar, A Aleman, and Luc Vanhees (2008). "Physical activity and enhanced fitness to improve cognitive function in older people without known cognitive impairment". In: *Cochrane database of systematic reviews* 2.

Anonymous (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *Submitted to International Conference on Learning*

Representations. under review. URL: https://openreview.net/forum?id=YicbFdNTTy.

Baloglu, Ulas Baran, Muhammed Talo, Özal Yildirim, Ru San Tan, and U. Rajendra Acharya (2019). "Classification of myocardial infarction with multi-lead ECG signals and deep CNN". In: *Pattern Recognit. Lett.*

Barness, Lewis A, John M Opitz, and Enid Gilbert-Barness (2007). "Obesity: genetic, molecular, and environmental aspects". In: *American journal of medical genetics part A* 143.24, pp. 3016–3034.

Bashyam, Vishnu M, Guray Erus, Jimit Doshi, Mohamad Habes, Ilya Nasralah, Monica Truelove-Hill, Dhivya Srinivasan, Liz Mamourian, Raymond Pomponio, Yong Fan, et al. (2020). "MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide". In: *Brain* 143.7, pp. 2312–2324.

Bayramoglu, Neslihan, Juho Kannala, and Janne Heikkilä (2016). "Deep learning for magnification independent breast cancer histopathology image classification". In: *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1, pp. 289–300. ISSN: 0035-9246.

Bergstra, James and Yoshua Bengio (2012). "Random Search for Hyper-Parameter Optimization". In: *The Journal of Machine Learning Research*.

Bietti, Alberto, Grégoire Mialon, Dexiong Chen, and Julien Mairal (2019). "A Kernel Perspective for Regularizing Deep Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32. ISSN: 0885-6125.

Brendel, Wieland, Jonas Rauber, and Matthias Bethge (2018). "Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Bronstein, Michael M., Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst (2017). "Geometric Deep Learning: Going beyond Euclidean data". In: *IEEE Signal Process. Mag.*

Cao, Bokai, Hucheng Zhou, Guoqiang Li, and Philip S Yu (2016). "Multi-view machines". In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, pp. 427–436.

Caruana, Rich, Steve Lawrence, and C. Lee Giles (2000). "Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping". In: *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*.

Cheng, Chi-Tung, Tsung-Ying Ho, Tao-Yi Lee, Chih-Chen Chang, Ching-Cheng Chou, Chih-Chi Chen, I-Fang Chung, and Chien-Hung Liao (2019). "Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs". In: *European radiology* 29.10, pp. 5469–5477.

Cheng, Yu, Felix X. Yu, Rogério Schmidt Feris, Sanjiv Kumar, Alok N. Choudhary, and Shih-Fu Chang (2015). "An Exploration of Parameter Redundancy in Deep Networks with Circulant Projections". In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*.

Choi, Shing Wan, Timothy Shin-Heng Mak, and Paul F O'Reilly (2020). "Tutorial: a guide to performing polygenic risk score analyses". In: *Nature Protocols* 15.9, pp. 2759–2772.

Cichocki, Andrzej, Namgil Lee, Ivan V. Oseledets, Anh Huy Phan, Qibin Zhao, and Danilo P. Mandic (2016). "Tensor Networks for Dimensionality Reduction and Large-scale Optimization: Part 1 Low-Rank Tensor Decompositions". In: *Found. Trends Mach. Learn.*

Coffey, Donald S (1998). "Self-organization, complexity and chaos: the new biology for medicine". In: *Nature medicine* 4.8, pp. 882–885.

Cole, James H (2020). "Multi-modality neuroimaging brain-age in UK Biobank: relationship to biomedical, lifestyle and cognitive factors". In: *Neurobiology of Aging*.

Cole, James H, Robert Leech, David J Sharp, and Alzheimer's Disease Neuroimaging Initiative (2015). "Prediction of brain age suggests accelerated atrophy after traumatic brain injury". In: *Annals of neurology* 77.4, pp. 571–581. ISSN: 0364-5134.

Cole, James H, Riccardo E Marioni, Sarah E Harris, and Ian J Deary (2019). "Brain age and other bodily 'ages': implications for neuropsychiatry". In: *Molecular psychiatry* 24.2, p. 266. ISSN: 1476-5578.

Cole, James H, Rudra PK Poudel, Dimosthenis Tsagkrasoulis, Matthan WA Caan, Claire Steves, Tim D Spector, and Giovanni Montana (2017). "Predicting brain

age with deep learning from raw imaging data results in a reliable and heritable biomarker". In: *NeuroImage* 163, pp. 115–124. ISSN: 1053-8119.

Cole, James H, Joel Raffel, Tim Friede, Arman Eshaghi, Wallace J Brownlee, Declan Chard, Nicola De Stefano, Christian Enzinger, Lukas Pirpamer, Massimo Filippi, et al. (2020). "Longitudinal assessment of multiple sclerosis with the brain-age paradigm". In: *Annals of Neurology* 88.1, pp. 93–105.

Cole, James H, Stuart J Ritchie, Mark E Bastin, MC Valdés Hernández, S Muñoz Maniega, Natalie Royle, Janie Corley, Alison Pattie, Sarah E Harris, and Qian Zhang (2018). "Brain age predicts mortality". In: *Molecular psychiatry* 23.5, p. 1385. ISSN: 1476-5578.

Collins, Francis S (2004). "What we do and don't know about 'race', 'ethnicity', genetics and health at the dawn of the genome era". In: *Nature genetics* 36.11, S13–S15.

Consortium, 1000 Genomes Project et al. (2012). "An integrated map of genetic variation from 1,092 human genomes". In: *Nature* 491.7422, pp. 56–65.

Cybenko, George (1989). "Approximation by superpositions of a sigmoidal function". In: *Mathematics of control, signals and systems* 2.4, pp. 303–314.

Davey Smith, George and Shah Ebrahim (2003). "'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?" In: *International journal of epidemiology* 32.1, pp. 1–22. ISSN: 1464-3685.

De Leeuw, F-E, Jan Cees de Groot, Matthijs Oudkerk, JCM Witteman, A Hofman, J Van Gijn, and MMB Breteler (2002). "Hypertension and cerebral white matter lesions in a prospective cohort study". In: *Brain* 125.4, pp. 765–772.

Deary, Ian J, Janie Corley, Alan J Gow, Sarah E Harris, Lorna M Houlihan, Riccardo E Marioni, Lars Penke, Snorri B Rafnsson, and John M Starr (2009). "Age-associated cognitive decline". In: *British medical bulletin* 92.1, pp. 135–152.

Duchi, John, Elad Hazan, and Yoram Singer (2011). "Adaptive subgradient methods for online learning and stochastic optimization." In: *Journal of machine learning research* 12.7.

Dudbridge, Frank (2013). "Power and predictive accuracy of polygenic risk scores". In: *PLoS genetics* 9.3, e1003348.

— (2016). "Polygenic epidemiology". In: *Genetic epidemiology* 40.4, pp. 268–272.

Ehret, Georg B, Patricia B Munroe, Kenneth M Rice, Murielle Bochud, Andrew D Johnson, Daniel I Chasman, Albert V Smith, Martin D Tobin, Germaine C Verwoert, and Shih-Jen Hwang (2011). "Genetic variants in novel pathways

influence blood pressure and cardiovascular disease risk". In: *Nature* 478.7367, p. 103. ISSN: 1476-4687.

Elliott, Joshua, Barbara Bodinier, Tom A Bond, Marc Chadeau-Hyam, Evangelos Evangelou, Karel GM Moons, Abbas Dehghan, David C Muller, Paul Elliott, and Ioanna Tzoulaki (2020). "Predictive accuracy of a polygenic risk score–enhanced prediction model vs a clinical risk score for coronary artery disease". In: *Jama* 323.7, pp. 636–645.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun (2017). "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nat.*

Esteva, Andre, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean (2019). "A guide to deep learning in healthcare". In: *Nature medicine* 25.1, pp. 24–29.

Evangelou, Evangelos, Helen R Warren, David Mosen-Ansorena, Borbala Mifsud, Raha Pazoki, He Gao, Georgios Ntritsos, Niki Dimou, Claudia P Cabrera, and Ibrahim Karaman (2018). "Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits". In: *Nature genetics* 50.10, p. 1412. ISSN: 1546-1718.

Filippi, M, C Tortorella, M Rovaris, M Bozzali, F Possa, MP Sormani, G Iannucci, and G Comi (2000). "Changes in the normal appearing brain tissue and cognitive impairment in multiple sclerosis". In: *Journal of Neurology, Neurosurgery and Psychiatry* 68.2, pp. 157–161.

Franke, Katja, Christian Gaser, Brad Manor, and Vera Novak (2013). "Advanced BrainAGE in older adults with type 2 diabetes mellitus". In: *Frontiers in aging neuroscience* 5, p. 90.

Franke, Katja, Christian Gaser, Tessa J Roseboom, Matthias Schwab, and Susanne R de Rooij (2018). "Premature brain aging in humans exposed to maternal nutrient restriction during early gestation". In: *NeuroImage* 173, pp. 460–471. ISSN: 1053-8119.

Fukushima, Kunihiko and Sei Miyake (1982). "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition". In: *Competition and cooperation in neural nets*. Springer, pp. 267–285.

Gabriel, Stacey B, Stephen F Schaffner, Huy Nguyen, Jamie M Moore, Jessica Roy, Brendan Blumenstiel, John Higgins, Matthew DeFelice, Amy Lochner, Maura Faggart, et al. (2002). "The structure of haplotype blocks in the human genome". In: *Science* 296.5576, pp. 2225–2229.

Gaser, Christian, Katja Franke, Stefan Klöppel, Nikolaos Koutsouleris, Heinrich Sauer, and Alzheimer's Disease Neuroimaging Initiative (2013). "BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease". In: *PloS one* 8.6, e67346. ISSN: 1932-6203.

Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio (2016). *Deep learning*. Vol. 1. 2. MIT press Cambridge.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy (2015). "Explaining and Harnessing Adversarial Examples". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Gudbjartsson, Hákon and Samuel Patz (1995). "The Rician distribution of noisy MRI data". In: *Magnetic resonance in medicine*.

Gulshan, Varun, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. (2016). "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs". In: *Jama* 316.22, pp. 2402–2410.

Gunning-Dixon, Faith M, Ruben C Gur, Alexis C Perkins, Lee Schroeder, Travis Turner, Bruce I Turetsky, Robin M Chan, James W Loughead, David C Alsop, and Joseph Maldjian (2003). "Age-related differences in brain activation during emotional face processing". In: *Neurobiology of aging* 24.2, pp. 285–295. ISSN: 0197-4580.

Hartley, Tom, Colin Lever, Neil Burgess, and John O'Keefe (2014). "Space in the brain: how the hippocampal formation supports spatial cognition". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1635, p. 20120510. ISSN: 0962-8436.

Havaei, Mohammad, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle (2017). "Brain tumor segmentation with deep neural networks". In: *Medical image analysis* 35, pp. 18–31.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*.

Hebb, Donald Olding (1949). *The organization of behavior: a neuropsychological theory*. J. Wiley; Chapman and Hall.

Hein, Matthias, Maksym Andriushchenko, and Julian Bitterwolf (2019). "Why ReLU Networks Yield High-Confidence Predictions Far Away From the Training Data and How to Mitigate the Problem". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*.

Hemani, Gibran, Jie Zheng, Benjamin Elsworth, Kaitlin H Wade, Valeriia Haberland, Denis Baird, Charles Laurin, Stephen Burgess, Jack Bowden, and Ryan Langdon (2018). "The MR-Base platform supports systematic causal inference across the human phenome". In: *Elife* 7, e34408. ISSN: 2050-084X.

Hitchcock, Frank L (1927). "The expression of a tensor or a polyadic as a sum of products". In: *Journal of Mathematics and Physics* 6.1-4, pp. 164–189.

Hoche, Franziska, Xavier Guell, Mark G Vangel, Janet C Sherman, and Jeremy D Schmahmann (2018). "The cerebellar cognitive affective/Schmahmann syndrome scale". In: *Brain* 141.1, pp. 248–270. ISSN: 0006-8950.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Høgestøl, Einar A, Tobias Kaufmann, Gro O Nygaard, Mona K Beyer, Piotr Sowa, Jan E Nordvik, Knut Kolskår, Geneviève Richard, Ole A Andreassen, Hanne F Harbo, et al. (2019). "Cross-sectional and longitudinal MRI brain scans reveal accelerated brain aging in multiple sclerosis". In: *Frontiers in neurology* 10, p. 450.

Howard, George, Maciej Banach, Mary Cushman, David C Goff, Virginia J Howard, Daniel T Lackland, Jim McVay, James F Meschia, Paul Muntner, and Suzanne Oparil (2015). "Is blood pressure control for stroke prevention the correct goal? The lost opportunity of preventing hypertension". In: *Stroke* 46.6, pp. 1595–1600. ISSN: 0039-2499.

Ioffe, Sergey and Christian Szegedy (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*.

Islam, Jyoti and Yanqing Zhang (2017). "A novel deep learning based multi-class classification method for Alzheimer's disease detection using brain MRI data". In: *International Conference on Brain Informatics*. Springer, pp. 213–222.

IXI Dataset (n.d.). URL: `https://brain-development.org/ixi-dataset/`.

Jacobs, Heidi IL, David A Hopkins, Helen C Mayrhofer, Emiliano Bruner, Fred W van Leeuwen, Wijnand Raaijmakers, and Jeremy D Schmahmann (2017). "The

cerebellum in Alzheimer's disease: evaluating its role in cognitive decline". In: *Brain* 141.1, pp. 37–47. ISSN: 0006-8950.

Jakubovitz, Daniel and Raja Giryes (2018). "Improving DNN Robustness to Adversarial Attacks Using Jacobian Regularization". In: *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*.

Jenkinson, Mark, Peter Bannister, Michael Brady, and Stephen Smith (2002). "Improved optimization for the robust and accurate linear registration and motion correction of brain images". In: *Neuroimage* 17.2, pp. 825–841. ISSN: 1053-8119.

Jernigan, Terry L, Doris A Trauner, John R Hesselink, and Paula A Tallal (1991). "Maturation of human cerebrum observed in vivo during adolescence". In: *Brain* 114.5, pp. 2037–2049.

Jurado, María Beatriz and Mónica Rosselli (2007). "The elusive nature of executive functions: a review of our current understanding". In: *Neuropsychology review* 17.3, pp. 213–233.

Kaczkurkin, Antonia N, Armin Raznahan, and Theodore D Satterthwaite (2019). "Sex differences in the developing brain: insights from multimodal neuroimaging". In: *Neuropsychopharmacology* 44.1, pp. 71–85.

Kaufmann, Tobias, Dennis van der Meer, Nhat Trung Doan, Emanuel Schwarz, Martina J Lund, Ingrid Agartz, Dag Alnæs, Deanna M Barch, Ramona Baur-Streubel, and Alessandro Bertolino (2019). "Common brain disorders are associated with heritable patterns of apparent aging of the brain". In: *Nature Neuroscience* 22.10, pp. 1617–1623. ISSN: 1546-1726.

Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres (2018). "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kodl, Christopher T and Elizabeth R Seaquist (2008). "Cognitive dysfunction and diabetes mellitus". In: *Endocrine reviews* 29.4, pp. 494–511.

Kolbeinsson, Arinbjörn, Sarah Filippi, Yannis Panagakis, Paul M Matthews, Paul Elliott, Abbas Dehghan, and Ioanna Tzoulaki (2020). "Accelerated MRI-predicted brain ageing and its associations with cardiometabolic and brain disorders". In: *Scientific Reports* 10.1, pp. 1–9.

Kolbeinsson, Arinbjörn, Jean Kossaifi, Yannis Panagakis, Adrian Bulat, Animashree Anandkumar, Ioanna Tzoulaki, and Paul M Matthews (2021). *Tensor dropout for robust learning*.

Kolda, Tamara G and Brett W Bader (2009). "Tensor decompositions and applications". In: *SIAM review* 51.3, pp. 455–500.

Kolenic, Marian, Katja Franke, Jaroslav Hlinka, Martin Matejka, Jana Capkova, Zdenka Pausova, Rudolf Uher, Martin Alda, Filip Spaniel, and Tomas Hajek (2018). "Obesity, dyslipidemia and brain age in first-episode psychosis". In: *Journal of psychiatric research* 99, pp. 151–158. ISSN: 0022-3956.

Kondor, Risi and Shubhendu Trivedi (2018). "On the Generalization of Equivariance and Convolution in Neural Networks to the Action of Compact Groups". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*.

Korolev, Sergey, Amir Safiullin, Mikhail Belyaev, and Yulia Dodonova (2017). "Residual and plain convolutional neural networks for 3D brain MRI classification". In: *14th IEEE International Symposium on Biomedical Imaging, ISBI 2017, Melbourne, Australia, April 18-21, 2017*.

Kossaifi, Jean, Adrian Bulat, Yannis Panagakis, and Maja Pantic (2019). "Efficient N-Dimensional Convolutions via Higher-Order Factorization". In: *CoRR* abs/1906.06196.

Kossaifi, Jean, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic (2019). "T-Net: Parametrizing Fully Convolutional Nets With a Single High-Order Tensor". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*.

Kossaifi, Jean, Aran Khanna, Zachary Lipton, Tommaso Furlanello, and Anima Anandkumar (2017). "Tensor Contraction Layers for Parsimonious Deep Nets". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*.

Kossaifi, Jean, Zachary C. Lipton, Arinbjörn Kolbeinsson, Aran Khanna, Tommaso Furlanello, and Anima Anandkumar (2020). In: *Journal of Machine Learning Research* 21.123, pp. 1–21.

Kossaifi, Jean, Yannis Panagakis, Anima Anandkumar, and Maja Pantic (2019). "Tensorly: Tensor learning in python". In: *The Journal of Machine Learning Research* 20.1, pp. 925–930. ISSN: 1532-4435.

Kovalev, Vassili A, Frithjof Kruggel, and D Yves von Cramon (2003). "Gender and age effects in structural brain asymmetry as measured by MRI texture analysis". In: *NeuroImage* 19.3, pp. 895–905. ISSN: 1053-8119.

Krizhevsky, Alex and Geoffrey Hinton (2009). *Learning multiple layers of features from tiny images*. Tech. rep. Citeseer.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *NeurIPS*.

Krogh, Anders and John A. Hertz (1991). "A Simple Weight Decay Can Improve Generalization". In: *Advances in Neural Information Processing Systems 4, [NIPS Conference, Denver, Colorado, USA, December 2-5, 1991]*.

Krysko, Kristen M, Roland G Henry, Bruce AC Cree, Jue Lin, UCSF MS-EPIC Team, Stacy Caillier, Adam Santaniello, Chao Zhao, Refujia Gomez, and Carolyn Bevan (2019). "Telomere length is associated with disability progression in multiple sclerosis". In: *Annals of neurology*. ISSN: 0364-5134.

Kunkle, Brian W, Benjamin Grenier-Boley, Rebecca Sims, Joshua C Bis, Vincent Damotte, Adam C Naj, Anne Boland, Maria Vronskaya, Sven J Van Der Lee, Alexandre Amlie-Wolf, et al. (2019). "Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing". In: *Nature genetics* 51.3, pp. 414–430.

Laksshman, Sundaram, Rajendra Rana Bhat, Vivek Viswanath, and Xiaolin Li (2017). "DeepBipolar: Identifying genomic mutations for bipolar disorder via deep learning". In: *Human mutation* 38.9, pp. 1217–1224.

Lambert, Jean-Charles, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, and Gary W Beecham (2013). "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease". In: *Nature genetics* 45.12, p. 1452. ISSN: 1546-1718.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *nature* 521.7553, p. 436.

LeCun, Yann, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel (1989). "Backpropagation applied to handwritten zip code recognition". In: *Neural computation* 1.4, pp. 541–551.

Lee, Chen-Yu, Patrick W. Gallagher, and Zhuowen Tu (2016). "Generalizing Pooling Functions in Convolutional Neural Networks: Mixed, Gated, and Tree". In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*.

Lemaitre, Herve, Aaron L Goldman, Fabio Sambataro, Beth A Verchinski, Andreas Meyer-Lindenberg, Daniel R Weinberger, and Venkata S Mattay (2012). "Normal age-related brain morphometric changes: nonuniformity across cortical

thickness, surface area and gray matter volume?" In: *Neurobiology of aging* 33.3, 617. e1–617. e9. ISSN: 0197-4580.

Li, Chunyuan, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski (2018). "Measuring the Intrinsic Dimension of Objective Landscapes". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Li, Xiaomeng, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng (2018). "H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes". In: *IEEE Trans. Medical Imaging*.

Linnainmaa, Seppo (1970). "The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors". In: *Master's Thesis (in Finnish), Univ. Helsinki*, pp. 6–7.

Litjens, Geert, Clara I Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen-Van De Kaa, Peter Bult, Bram Van Ginneken, and Jeroen Van Der Laak (2016). "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis". In: *Scientific reports* 6, p. 26286.

Liu, Ningning, Emmanuel Dellandréa, Liming Chen, Chao Zhu, Yu Zhang, Charles-Edmond Bichot, StéPhane Bres, and Bruno Tellez (2013). "Multimodal recognition of visual concepts using histograms of textual concepts and selective weighted late fusion scheme". In: *Computer Vision and Image Understanding* 117.5, pp. 493–512.

Locke, Adam E, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. (2015). "Genetic studies of body mass index yield new insights for obesity biology". In: *Nature* 518.7538, pp. 197–206.

Lundberg, Scott M. and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*.

Mak, Timothy Shin Heng, Robert Milan Porsch, Shing Wan Choi, Xueya Zhou, and Pak Chung Sham (2017). "Polygenic scores via penalized regression on summary statistics". In: *Genetic epidemiology* 41.6, pp. 469–480.

Massion, Pierre P, Sanja Antic, Sarim Ather, Carlos Arteta, Jan Brabec, Heidi Chen, Jerome Declerck, David Dufek, William Hickes, Timor Kadir, et al. (2020). "Assessing the Accuracy of a Deep Learning Method to Risk Stratify Indeter-

minate Pulmonary Nodules". In: *American Journal of Respiratory and Critical Care Medicine* ja.

Matthews, Fiona E, Antony Arthur, Linda E Barnes, John Bond, Carol Jagger, Louise Robinson, Carol Brayne, Medical Research Council Cognitive Function, and Ageing Collaboration (2013). "A two-decade comparison of prevalence of dementia in individuals aged 65 years and older from three geographical areas of England: results of the Cognitive Function and Ageing Study I and II". In: *The Lancet* 382.9902, pp. 1405–1412. ISSN: 0140-6736.

Mavaddat, Nasim, Kyriaki Michailidou, Joe Dennis, Michael Lush, Laura Fachal, Andrew Lee, Jonathan P Tyrer, Ting-Huei Chen, Qin Wang, Manjeet K Bolla, et al. (2019). "Polygenic risk scores for prediction of breast cancer and breast cancer subtypes". In: *The American Journal of Human Genetics* 104.1, pp. 21–34.

McInnes, Leland and John Healy (2018). "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *arXiv preprint arXiv:1802.03426*.

McMahan, H. Brendan and Matthew J. Streeter (2010). "Adaptive Bound Optimization for Online Convex Optimization". In: *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*.

Meng, Lei, Xin-Yu Li, Liang Shen, and Hong-Fang Ji (2020). "Type 2 Diabetes Mellitus Drugs for Alzheimer's Disease: Current Evidence and Therapeutic Opportunities". In: *Trends in Molecular Medicine*.

Menon, Vinod and Lucina Q Uddin (2010). "Saliency, switching, attention and control: a network model of insula function". In: *Brain Structure and Function* 214.5-6, pp. 655–667. ISSN: 1863-2653.

Millard, Louise AC, Neil M Davies, Tom R Gaunt, George Davey Smith, and Kate Tilling (2017). "Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank". In: *International journal of epidemiology*.

Miller, Karla L, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, and Jesper LR Andersson (2016). "Multimodal population brain imaging in the UK Biobank prospective epidemiological study". In: *Nature neuroscience* 19.11, p. 1523. ISSN: 1546-1726.

Min, Seonwoo, Byunghan Lee, and Sungroh Yoon (2017). "Deep learning in bioinformatics". In: *Briefings Bioinform.*

Mitchell, Tom M et al. (1997). "Machine learning". In:

Morris, Andrew P, Benjamin F Voight, Tanya M Teslovich, Teresa Ferreira, Ayellet V Segre, Valgerdur Steinthorsdottir, Rona J Strawbridge, Hassan Khan, Harald Grallert, and Anubha Mahajan (2012). "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes". In: *Nature genetics* 44.9, p. 981. ISSN: 1546-1718.

Muller, Majon, Sigurdur Sigurdsson, Olafur Kjartansson, Thor Aspelund, Oscar L Lopez, Palmi V Jonnson, Tamara B Harris, Mark Van Buchem, Vilmundur Gudnason, Lenore J Launer, et al. (2014). "Joint effect of mid-and late-life blood pressure on the brain: the AGES-Reykjavik study". In: *Neurology* 82.24, pp. 2187–2195.

Nature Editorial (Nov. 2016). "The power of big data must be harnessed for medical progress". In: *Nature* 539.7630, pp. 467–468. DOI: 10.1038/539467b. URL: https://doi.org/10.1038/539467b.

Nikpay, Majid, Anuj Goel, Hong-Hee Won, Leanne M Hall, Christina Willenborg, Stavroula Kanoni, Danish Saleheen, Theodosios Kyriakou, Christopher P Nelson, Jemma C Hopewell, et al. (2015). "A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease". In: *Nature genetics* 47.10, p. 1121.

Ning, Kaida, Lu Zhao, Will Matloff, Fengzhu Sun, and Arthur W Toga (2020). "Association of relative brain age with tobacco smoking, alcohol consumption, and genetic variants". In: *Scientific reports* 10.1, pp. 1–10.

Nojavanasghari, Behnaz, Deepak Gopinath, Jayanth Koushik, Tadas Baltrusaitis, and Louis-Philippe Morency (2016). "Deep multimodal fusion for persuasiveness prediction". In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016*.

Norcliffe-Brown, Will, Stathis Vafeias, and Sarah Parisot (2018). "Learning Conditioned Graph Structures for Interpretable Visual Question Answering". In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*.

Novikov, Alexander, Mikhail Trofimov, and Ivan V. Oseledets (2017). "Exponential Machines". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*.

Nowlan, Steven J. and Geoffrey E. Hinton (1992). "Simplifying Neural Networks by Soft Weight-Sharing". In: *Neural Comput.*

Oord, Aäron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu

(2016). "WaveNet: A Generative Model for Raw Audio". In: *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*.

Oseledets, Ivan V (2011). "Tensor-train decomposition". In: *SIAM Journal on Scientific Computing* 33.5, pp. 2295–2317.

Paszke, Adam, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer (2017). "Automatic differentiation in pytorch". In:

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*.

Pawlowski, Nick and Ben Glocker (2019). "Is Texture Predictive for Age and Sex in Brain MRI?" In: *arXiv preprint arXiv:1907.10961*.

Pawlowski, Nick, Sofia Ira Ktena, Matthew C. H. Lee, Bernhard Kainz, Daniel Rueckert, Ben Glocker, and Martin Rajchl (2017). "DLTK: State of the Art Reference Implementations for Deep Learning on Medical Images". In: *arXiv preprint arXiv:1711.06853*.

Peng, Yuxin, Jinwei Qi, Xin Huang, and Yuxin Yuan (2017). "CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network". In: *IEEE Transactions on Multimedia* 20.2, pp. 405–420.

Pérez-Rosas, Verónica, Rada Mihalcea, and Louis-Philippe Morency (2013). "Utterance-level multimodal sentiment analysis". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 973–982.

Pfefferbaum, Adolf, Daniel H Mathalon, Edith V Sullivan, Jody M Rawles, Robert B Zipursky, and Kelvin O Lim (1994). "A quantitative magnetic resonance imaging study of changes in brain morphology from infancy to late adulthood". In: *Archives of neurology* 51.9, pp. 874–887.

Poplin, Ryan, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, et al. (2018). "A universal SNP and small-indel variant caller using deep neural networks". In: *Nature biotechnology* 36.10, pp. 983–987.

Poplin, Ryan, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster (2018). "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning". In: *Nature Biomedical Engineering* 2.3, p. 158.

Qiu, Chengxuan and Laura Fratiglioni (2015). "A major role for cardiovascular burden in age-related cognitive decline". In: *Nature Reviews Cardiology* 12.5, p. 267.

Raji, Cyrus A, April J Ho, Neelroop N Parikshak, James T Becker, Oscar L Lopez, Lewis H Kuller, Xue Hua, Alex D Leow, Arthur W Toga, and Paul M Thompson (2010). "Brain structure and obesity". In: *Human brain mapping* 31.3, pp. 353–364.

Rajkomar, Alvin, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. (2018). "Scalable and accurate deep learning with electronic health records". In: *NPJ Digital Medicine* 1.1, p. 18.

Rauber, Jonas, Wieland Brendel, and Matthias Bethge (2017). "Foolbox v0.8.0: A Python toolbox to benchmark the robustness of machine learning models". In: *arXiv preprint arXiv:1707.04131*.

Raz, Naftali, Ulman Lindenberger, Karen M Rodrigue, Kristen M Kennedy, Denise Head, Adrienne Williamson, Cheryl Dahle, Denis Gerstorf, and James D Acker (2005). "Regional brain changes in aging healthy adults: general trends, individual differences and modifiers". In: *Cerebral cortex* 15.11, pp. 1676–1689. ISSN: 1047-3211.

Rendle, Steffen (2010). "Factorization machines". In: *2010 IEEE International Conference on Data Mining*. IEEE, pp. 995–1000.

Romero, Adriana, Pierre Luc Carrier, Akram Erraqabi, Tristan Sylvain, Alex Auvolat, Etienne Dejoie, Marc-André Legault, Marie-Pierre Dubé, Julie G. Hussin, and Yoshua Bengio (2017). "Diet Networks: Thin Parameters for Fat Genomics". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*.

Rosenblatt, Frank (1958). "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6, p. 386.

Rowe, John W and Robert L Kahn (1987). "Human aging: usual and successful". In: *Science* 237.4811, pp. 143–149. ISSN: 0036-8075.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986). "Learning representations by back-propagating errors". In: *nature* 323.6088, pp. 533–536.

Sarraf, Saman and Ghassem Tofighi (2016). "Deep learning-based pipeline to recognize Alzheimer's disease using fMRI data". In: *2016 Future Technologies Conference (FTC)*. IEEE, pp. 816–820.

Scardapane, Simone, Danilo Comminiello, Amir Hussain, and Aurelio Uncini (2017). "Group sparse regularization for deep neural networks". In: *Neurocomputing* 241, pp. 81–89.

Schrepf, Andrew, Chelsea M Kaplan, Eric Ichesco, Tony Larkin, Steven E Harte, Richard E Harris, Alison D Murray, Gordon D Waiter, Daniel J Clauw, and Neil Basu (2018). "A multi-modal MRI study of the central response to inflammation in rheumatoid arthritis". In: *Nature communications* 9.1, pp. 1–11.

Senior, Andrew W., Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Zídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis (2020). "Improved protein structure prediction using potentials from deep learning". In: *Nat.*

Shafer, Douglas S. (1995). "Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering (Steven H. Strogatz)". In: *SIAM Rev.*

Smith, S, FA Almagro, and K Miller (2017). *UK Biobank Brain Imaging Documentation, Version 1.7*.

Smith, Stephen M, Lloyd T Elliott, Fidel Alfaro-Almagro, Paul McCarthy, Thomas E Nichols, Gwenaëlle Douaud, and Karla L Miller (2020). "Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations". In: *Elife* 9, e52677.

Smith, Stephen M, Diego Vidaurre, Fidel Alfaro-Almagro, Thomas E Nichols, and Karla L Miller (2019). "Estimation of brain age delta from brain imaging". In: *NeuroImage*. ISSN: 1053-8119.

Soltaninejad, Mohammadreza, Guang Yang, Tryphon Lambrou, Nigel M. Allinson, Timothy L. Jones, Thomas R. Barrick, Franklyn A. Howe, and Xujiong Ye (2018). "Supervised Learning based Multimodal MRI Brain Tumour Segmentation using Texture Features from Supervoxels". In: *Comput. Methods Programs Biomed.*

Sowell, Elizabeth R, Bradley S Peterson, Paul M Thompson, Suzanne E Welcome, Amy L Henkenius, and Arthur W Toga (2003). "Mapping cortical change across the human life span". In: *Nature neuroscience* 6.3, p. 309. ISSN: 1546-1726.

Sowell, Elizabeth R, Paul M Thompson, Colin J Holmes, Rajneesh Batth, Terry L Jernigan, and Arthur W Toga (1999). "Localizing age-related changes in brain structure between childhood and adolescence using statistical parametric mapping". In: *Neuroimage* 9.6, pp. 587–597.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: a simple way to prevent neural networks from overfitting". In: *JMLR* 15.1, pp. 1929–1958.

Stern, Yaakov (2012). "Cognitive reserve in ageing and Alzheimer's disease". In: *The Lancet Neurology* 11.11, pp. 1006–1012. ISSN: 1474-4422.

Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. (2015a). "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age". In: *PLoS medicine* 12.3.

Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, and Martin Landray (2015b). "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age". In: *PLoS medicine* 12.3, e1001779. ISSN: 1549-1676.

Sui, Jing, Shile Qi, Theo GM van Erp, Juan Bustillo, Rongtao Jiang, Dongdong Lin, Jessica A Turner, Eswar Damaraju, Andrew R Mayer, Yue Cui, et al. (2018). "Multimodal neuromarkers in schizophrenia via cognition-guided MRI fusion". In: *Nature communications* 9.1, pp. 1–14.

Sutskever, Ilya, James Martens, George E. Dahl, and Geoffrey E. Hinton (2013). "On the importance of initialization and momentum in deep learning". In: *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le (2014). "Sequence to Sequence Learning with Neural Networks". In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*.

Suzuki, Hideaki, Ashwin V Venkataraman, Wenjia Bai, Florian Guitton, Yike Guo, Abbas Dehghan, and Paul M Matthews (2019). "Associations of regional brain structural differences with aging, modifiable risk factors for dementia, and cognitive performance". In: *JAMA network open* 2.12, e1917257–e1917257.

Swan, Gary E, Charles DeCarli, BL Miller, T Reed, PA Wolf, LM Jack, and D Carmelli (1998). "Association of midlife blood pressure to late-life cognitive decline and brain morphology". In: *Neurology* 51.4, pp. 986–993.

Tai, Cheng, Tong Xiao, Xiaogang Wang, and Weinan E (2016). "Convolutional neural networks with low-rank regularization". In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Tan, Mingxing and Quoc V. Le (2019). "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*.

Tieleman, Tijmen and Geoffrey Hinton (2012). "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude". In: *COURSERA: Neural networks for machine learning* 4.2, pp. 26–31.

Tomasi, Dardo and Nora D Volkow (2012). "Aging and functional brain networks". In: *Molecular psychiatry* 17.5, pp. 549–558.

Tucker, Ledyard R (1966). "Some mathematical notes on three-mode factor analysis". In: *Psychometrika* 31.3, pp. 279–311.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*.

Veličković, Petar, Duo Wang, Nicholas D Lane, and Pietro Liò (2016). "X-CNN: Cross-modal convolutional neural networks for sparse datasets". In: *2016 IEEE symposium series on computational intelligence (SSCI)*. IEEE, pp. 1–8.

Wallace, Chris (2020). "Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses". In: *PLoS Genetics* 16.4, e1008720.

Walters, Kevin, Angela Cox, and Hannuun Yaacob (2019). "Using GWAS top hits to inform priors in Bayesian fine-mapping association studies". In: *Genetic epidemiology* 43.6, pp. 675–689.

Walther, Katrin, Alex C Birdsill, Elizabeth L Glisky, and Lee Ryan (2010). "Structural brain differences and cognitive functioning related to body mass index in older females". In: *Human brain mapping* 31.7, pp. 1052–1064.

Wan, Li, Matthew D. Zeiler, Sixin Zhang, Yann LeCun, and Rob Fergus (2013). "Regularization of Neural Networks using DropConnect". In: *Proceedings of the*

30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013.

Wang, Liqun, H-Ming Lai, Gareth J Barker, David H Miller, and Paul S Tofts (1998). "Correction for variations in MRI scanner sensitivity in brain studies with histogram matching". In: *Magnetic resonance in medicine* 39.2, pp. 322–327.

Ward, Michael A, Cynthia M Carlsson, Mehul A Trivedi, Mark A Sager, and Sterling C Johnson (2005). "The effect of body mass index on global brain volume in middle-aged adults: a cross sectional study". In: *BMC neurology* 5.1, p. 23.

Wray, Naomi R, Sang Hong Lee, Divya Mehta, Anna AE Vinkhuyzen, Frank Dudbridge, and Christel M Middeldorp (2014). "Research review: polygenic methods and their application to psychiatric traits". In: *Journal of Child Psychology and Psychiatry* 55.10, pp. 1068–1087.

Yang, Ge and Samuel Schoenholz (2017). "Mean field residual networks: On the edge of chaos". In: *Advances in neural information processing systems*, pp. 7103–7114.

Yang, Jian, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. (2010). "Common SNPs explain a large proportion of the heritability for human height". In: *Nature genetics* 42.7, pp. 565–569.

Yu, Xiyu, Tongliang Liu, Xinchao Wang, and Dacheng Tao (2017). "On Compressing Deep Models by Low Rank and Sparse Decomposition". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*.

Zhang, Yuchen, Jason D. Lee, and Michael I. Jordan (2016). "L1-regularized Neural Networks are Improperly Learnable in Polynomial Time". In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*.

Zhuang, Zhong, Xiaotong Shen, and Wei Pan (2019). "A simple convolutional neural network for prediction of enhancer-promoter interactions with DNA sequence data". In: *Bioinform.*

Zoph, Barret and Quoc V. Le (2017). "Neural Architecture Search with Reinforcement Learning". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.