Challenges for machine learning in clinical translation of big data imaging studies

Nicola K Dinsdale[1,2*], Emma Bluemke[3], Vaanathi Sundaresan[1,4], Mark Jenkinson[1,5,6], Stephen M Smith[1], Ana IL Namburete[2]

[1] Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, UK

[2] Oxford Machine Learning in NeuroImaging Lab, OMNI, Department of Computer Science, University of Oxford, UK

[3] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK

[4] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Charlestown, MA, United States

[5] Australian Institute for Machine Learning (AIML), School of Computer Science, University of Adelaide, Adelaide, Australia

[6] South Australian Health and Medical Research Institute (SAHMRI), North Terrace, Adelaide, Australia.

[*] Corresponding Author: nicola.dinsdale@cs.ox.ac.uk

Abstract

Combining deep learning image analysis methods and large-scale imaging datasets offers many opportunities to neuroscience imaging and epidemiology. However, despite these opportunities and the success of deep learning when applied to a range of neuroimaging tasks and domains, significant barriers continue to limit the impact of large-scale datasets and analysis tools. Here, we examine the main challenges and the approaches that have been explored to overcome them. We focus on issues relating to data availability, interpretability, evaluation, and logistical challenges, and discuss the problems that still need to be tackled to enable the success of 'big data' deep learning approaches beyond research.

1. Introduction

The majority of neuroimaging datasets have been limited to small-scale low-N collections, typically focusing on a specific research question or clinical population of interest. However, large-scale 'big data' collections of a wide range of subjects have begun to be collated, many of which are openly available to researchers. This means that if the acquisition protocol, demographic and non-imaging data meet the requirements of a given study, novel research can be completed without acquiring new scans. Sharing these large-scale datasets has had many benefits: they enable exploration of new research questions, and reproducible, rapid methodological prototyping.

Existing large-scale datasets have been curated to explore different research questions, with varying numbers of subjects and imaging sites across studies. For instance, if the research question were about lifespan and ageing, datasets to consider would include UK Biobank (Sudlow, et al. 2015) and CamCAN (Taylor, et al. 2017). Similarly, if considering early development, available datasets include the Developing HCP (dHCP) (Hughes, et al. 2017), and the Adolescent Brain Cognitive Development (ABCD) (Marek, et al. 2019); for research on young adults, one could consider HCP Young Adult (Van Essen, et al. 2013). Datasets also exist that explore specific clinical groups, such as Alzheimer's disease (ADNI (Jack, et al. 2008)) schizophrenia, and bipolar disorder (CANDI (Frazier, et al. 2008)). These datasets allow exploration of questions that would not be possible with traditional small-scale studies (e.g. with N<100), which will not sufficiently represent variation within the population of interest. Large-scale studies have also enabled the characteristation of potential subtypes within patient samples – for example, (Young, et al. 2018) demonstrated heterogeneity and subtypes in Alzheimer's related atrophy patterns using data from ADNI.

UK Biobank (Sudlow, et al. 2015), the largest of these studies, aims to collect brain imaging data from 100,000 volunteers, including 6 MRI modalities, to study structure, function, and connectivity. It contains a diverse range of lifestyle, genetic, and health measures, allowing

researchers to create models of population ageing and to explore how genetic and environmental factors interact with ageing and disease. For instance, hippocampal atrophy is a well-validated biomarker for Alzheimer's disease, so using the UK Biobank, a nomogram of hippocampal volume with normal ageing has been created (Nobis, et al. 2019), illustrating the progression with age, and percentiles of expected volume across for the healthy population, as a reference.

Due to the growth in size of these datasets, sophisticated deep learning models are finally a practical option for neuroimaging analysis, enabling exploration of new questions in a data-driven manner. Powered by their ability to learn complex, non-linear relationships and patterns from data, deep learning methods have been applied to a wide range of applications, finding success in previously unsolved problems. However, this success has been limited to specific tasks and data domains. Challenges remain for applying deep learning models to the clinical domain, which currently limit the impact that big datasets such as the UK Biobank have on patient care. Work must be undertaken to allow models to extend beyond the research domain. Recent developments in deep learning have begun to tackle the problems faced, but further developments are needed. Here, we discuss the challenges faced, the current approaches being developed to mitigate them, and the barriers that remain, including the challenges of data availability, interpretability and model evaluation, and logistical challenges such as data privacy.

2. Deep Learning Background

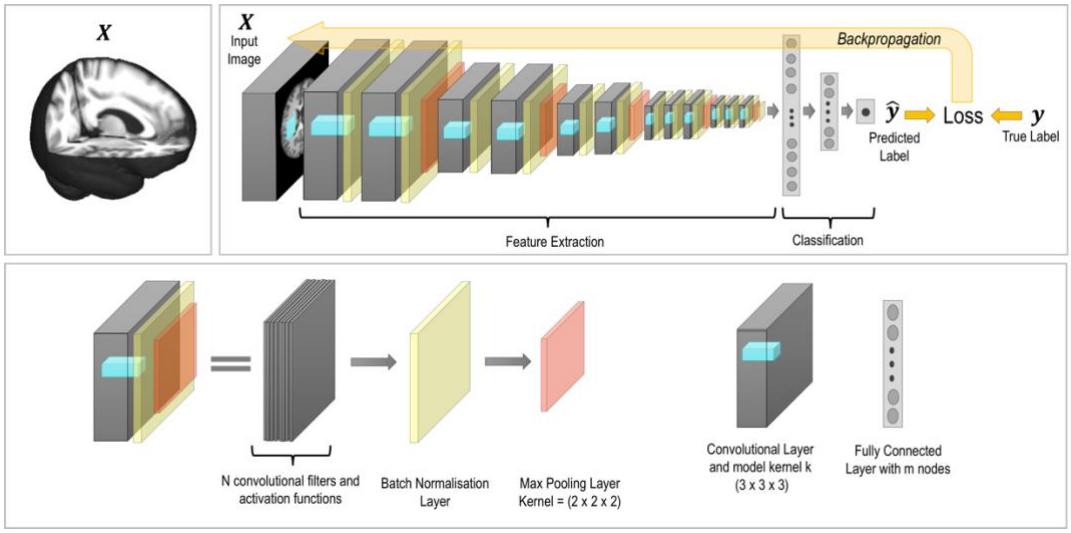

Figure 1: An example network architecture for a convolutional neural network (CNN) for a classification or regression task. The lower panel describes the building blocks used to form the network shown in the upper panel.

To understand the challenges for clinical translatability of deep learning methods, we first require a brief overview of how these methods approach problems - for a more detailed introduction see (LeCun, Bengio and Hinton 2015). We will only consider convolutional neural networks (CNNs), which form the vast majority of deep learning methods currently applied in medical imaging, an example architecture of which is shown in Fig. 1. The majority are *supervised* approaches (LeCun, Bengio and Hinton 2015) meaning that to explore the research question, we need a dataset of images, X, and a set of known true labels, y, for the task in question. This requires an understanding of both the information that we expect to be encoded within the images, and of which questions are of interest (defined as *domain knowledge*). An example for the variables could be a structural brain scan (X) with the label being disease prognosis. The task is then to design a neural network architecture capable of mapping from X to y through learning a highly non-linear mapping function $f(X, y; W)$, where W are the trainable weights of the neural network.

The choice of architecture is highly influenced by factors such as the task being explored, the quantity of data and the computational power available. Nevertheless, most networks are constructed from the same basic building blocks. First, *convolutional filters* which learn features of interest from the data (*feature extraction*). They contain the weights and biases to be learned during the optimisation process. Stacks of these layers are placed at different spatial resolutions for a range of different features to be extracted at each level of abstraction. This hierarchical feature extraction allows a rich understanding of the input data. During the *forward pass* of the training procedure, each filter is convolved across the width and height of the input volume. The exact nature of the features is learned through a network optimisation procedure that updates the filter weights, to find features that are useful contributors to the overall goal of predicting y.

Next are the *activation functions* which play a fundamental role in model training by applying nonlinear transformations to the learned features. This non-linearity provides a distinct edge to CNNs, allowing them to learn the complex non-linear relationships (or mapping) between the input and the output. Commonly used activation functions include rectified linear units (ReLU, e.g., zeroing negative values and keeping positive ones unchanged) and sigmoid (e.g., squashing large values down to a predefined upper bound, typically between 0 and 1). Due to the CNN's sequential data flow, the features at a given depth are a non-linear combination of the previous features and the network parameters, whose values are learned during network training. Without activation functions, CNNs would only be able to train linear models.

Networks then learn features at different spatial resolutions through the inclusion of *pooling blocks*. Pooling provides a basic invariance to rotations and translations, and has been demonstrated to improve the object detection capability of convolutional networks. The final key components of neural networks are *fully connected layers* - essential to many classification or regression architectures – which are normally placed at the end of a network and learn how to classify the extracted features.

By feeding the data through the network, we obtain an output prediction. To render these accurate, the weights of the network must be optimised through *back propagation*. To this end, we evaluate a *loss* or *cost function* which determines the error in the network prediction by comparing the prediction $\hat{y}$ and the true label y. The choice of loss function is task-dependent and plays a crucial role in the network performance.

Thus, we have an optimisation problem, the performance of which is highly dependent on two factors: first, the design decisions made about the network architecture and the loss function; second, the data available to train the network. Nearly all relevant techniques have been developed in computer vision, where very large datasets are available and easily curated, for instance by scraping the internet. In neuroimaging, data has to be labelled by a domain expert. This is one of many differences between neuroimaging and the computer vision field; many challenges are specific to working with neuroimaging data, especially when the aim is clinical translation.

3. Data Availability

For clinical translatability or for deep learning techniques to be applied to clinical research, data availability is a major limitation. Despite growth in the size of available datasets, the largest are still only of the order of tens of thousands, with a thousand images being commonly regarded as a large dataset. For many specific tasks, datasets exist only in the order of hundreds of subjects, due to factors including monetary and time costs of acquiring data, difficulties in sharing and/or pooling data across sites, and the fact that, for some conditions, insufficient patient numbers exist to create a dataset of any great size (Morid, Borjali and Del Fiol 2021). E.g., the frequently explored Brain Tumour Segmentation (BraTS) dataset (Menze, et al. 2014) only has data from 369 subjects available for training (2020 challenge data), in stark contrast to popular datasets from computer vision, such as ImageNet (Deng, et al. 2009)(1,281,167 training examples) and MNIST (LeCun, Bottou, et al. 1998)(60,000 training examples). Simply by considering dataset size, it is clear that we are likely to be underpowered for training neural networks: highly parameterised, deep neural networks are very

dependent on the amount of available training data (He, et al. 2020). With performance generally improving as the number of data points is increased, they are more affected by the amount of available training data than classical machine learning techniques, due to the need to learn the useful features as well as the (highly nonlinear) decision boundary (He, et al. 2020), and so techniques to overcome the lack of data are required, especially for clinical applications.

3.1 Maximising the impact of available data

There has, therefore, been an increasing focus on developing techniques to facilitate more effective use of available data. A commonly used technique from computer vision is the use of large natural image datasets (Raghu, et al. 2019), with ImageNet (Deng, et al. 2009) being the most popular, to *pre-train* the network. This involves training the weights on a related task with more available data, so the optimisation starts from an informed place, rather than a random initialisation. Clearly, this might be useful by considering the information learned by the network at the different stages (Olah, et al. 2018): the early layers learn features such as edges and simple textures, largely resembling Gabor filters, and are thus very general and applicable across different images, regardless of the target tasks (Yosinski, et al. 2014). The final layers learn features which are far more task- and dataset-specific. Therefore, we can take a network pre-trained on the large, canonical dataset, and use this to extract features which we then pass to a classifier, requiring only the final classifier layers to be trained, or, more commonly, the deeper layers can also be fine-tuned (re-trained) to the specific task. This requires less data, as not only are we starting the optimisation process from an informed point in the *parameter space*, but also the very earliest layers can often be frozen (kept at their value and not updated during training), greatly reducing the number of weights in the model that need to be optimised. This process is referred to as *transfer learning*; it is a step frequently used to allow networks to be trained with smaller amounts of training data. Transfer learning can be performed across data domain (dataset), task, or both, depending on the datasets available for pre-training, and so may enable us to train models on the clinical data of interest, and so explore clinical research questions directly – e.g. (Peng, et al. 2021) showed that by training on UK Biobank data and then finetuning the model on the target dataset, they significantly improved age prediction performance.

Although standard practice is to use the huge datasets of natural images for pre-training, natural images have very different characteristics from many medical images, so the features learned are not necessarily the most appropriate for the tasks being considered in neuroimaging (Raghu, et al. 2019). For instance, natural images are often stored as RGB 2D images (3 channel images), whereas MR images are encoded as greyscale (single-channel) 3D images. Also, in medical images, the location of structures could be informative, which is rarely true in natural images. Creating pre-trained networks for medical images has therefore been a focus, with Model Genesis (Zhou, et al. 2019) creating a flexible architecture trained to complete multiple tasks, extracting features which aim to generalise across medical imaging tasks. Similarly, some works pre-train on large datasets such as UK Biobank for tasks such as age or sex prediction, where obtaining labels is relatively trivial (Lu, et al. 2021) or on datasets for the same task with a dataset where more labels are available (Kushibar, et al. 2019). Again, the aim is to learn features from another task which are also useful for the task of interest - features that generalise across tasks, and information from a large dataset which helps us to understand a smaller clinical dataset.

Other studies utilise self-supervised approaches, such as *contrastive representation learning*, where general features of a dataset are learned, without labels, by teaching the model which data points are similar or different. These then act as the starting point for further model training on a smaller target dataset, rather than pre-training the model on a different dataset. An example approach is presented in (Chen, et al. 2020), where the data has been *augmented* (small transformations applied to increase the size of the dataset, discussed below); the network is then trained to encode both the original and the augmented images into the same location in the feature space using a contrastive loss function (Hadsell, Chopra and Lecun 2006) that learns features

describing the similarity between images. Different self-supervised methods and contrastive loss approaches have been developed and have begun to be applied in medical imaging (Zhang, et al. 2020) (Chaitanya, Erdil, et al. 2020), including for segmentation of MRI scans of the brain (Chen, et al. 2019).

3.2 Data Augmentation

CNNs, however, still ultimately require a reasonable amount of data (100s or 1000s) in the target data domain, as at least some of the network parameters must be fine-tuned to optimise the prediction performance for the specific dataset and task. Even though the amount of data required is likely to be reduced, the degree of reduction will be determined by the similarity between the proxy and target tasks (He, Girshick and Dollar 2019); the amount required may remain greater than is available. In this circumstance, data augmentation is often applied (Simard, Lecun and Denker 1998) artificially increasing the size and diversity of the training dataset by applying transformations, creating slightly perturbed versions of the data. Fundamentally, data augmentation enables us to artificially create a larger dataset which can be used to train the model, potentially enabling exploration directly with clinical data.

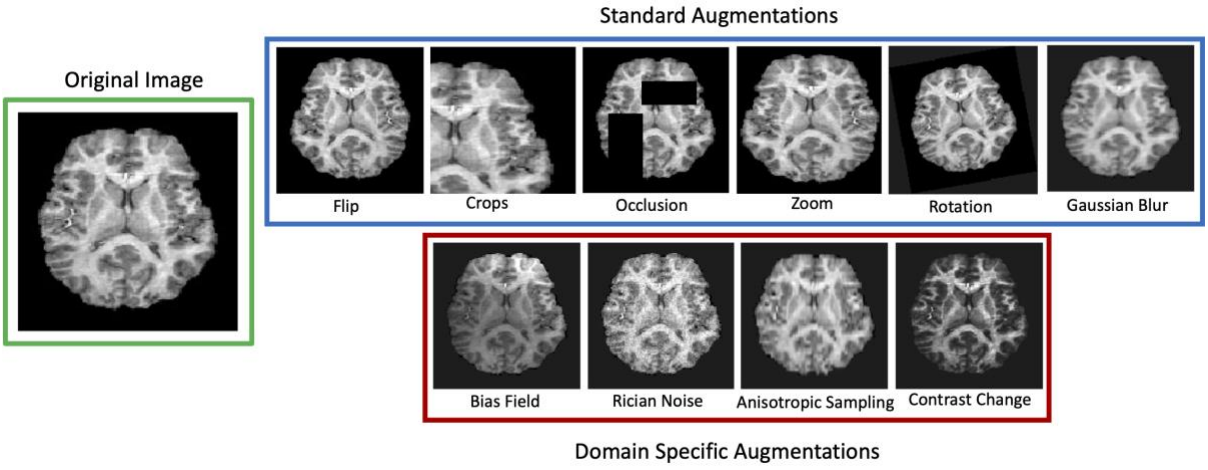

Figure 2: Example augmentations that might be applied to an MRI image. Standard augmentations come directly from computer vision approaches, and domain specific augmentations for neuroimaging focus on variation that would be likely to be seen in MR images.

Augmentations (Fig. 2) can take the form of basic transformations such as flips and rotations (Krizhevsky, Sutskever and Hinton 2012) (Simonyan and Zisserman 2015) as standardly applied in computer vision tasks, to more extreme examples such as Mixup (Zhang, et al. 2017) which merges images from different classes to form hybrid classes, or generative networks such as conditional Generative Adversarial Networks (GANs) - networks trained to generate simulated data (Mirza and Osindero 2014). While most deep learning studies apply data augmentation during training, some studies explore this for neuroimaging specifically: e.g., augmentation can be achieved through GANs being used to generate additional meaningful datapoints (Wu, et al. 2020), or registration to templates (Nguyen, et al. 2020), which generate biologically plausible transformations of the data. Similarly, they can be produced by identifying augmentations which are plausible across sites and scanners (Billot, Bocchetta, et al. 2020), such as applying bias field.

Existing literature suggests that performing augmentations, even transformations which create images beyond realistic variation (Billot, Bocchetta, et al. 2020), helps the network to generalise better to unseen data at test time. However, data augmentation must be used cautiously, so that the transformations applied do not change the validity of the label associated with the image. Consider, for instance, classifying Alzheimer's disease from structural MRI: the key indicator could be the atrophy of the hippocampus, so if any transformations are applied during the augmentation process that affect this region (e.g., local elastic deformations), it must be ensured that the level of atrophy is

not affected and, thus, the true label changed. Ensuring this requires high levels of specific domain knowledge and can limit the augmentations which can be applied.

3.3 Patch or Slice-based Sampling

Other approaches to solving the shortage of available training data focus on breaking the input data down into patches e.g. (Wachinger, Reuter and Klein 2018) or slices (where the data is 3D), with many studies treating MRI data as 2D inputs, where each slice is treated as a separate training sample, e.g., (Livne, et al. 2019). This approach can vastly increase the amount of available data and can be especially effective for segmentation tasks where we have voxel-level labels. However, fragmenting the image can lead to the loss of global information; when they can be implemented, fully 3D networks have in most cases provided better results (Kamnitsas, Ledig, et al. 2017). Patch-wise or slice-wise approaches cannot necessarily be applied to classification tasks; where a single label is provided for the whole image, it may not hold for a given patch or slice of the image (Khagi, Lee and Kwon 2018).

3.4 Differences between datasets or data domain shift

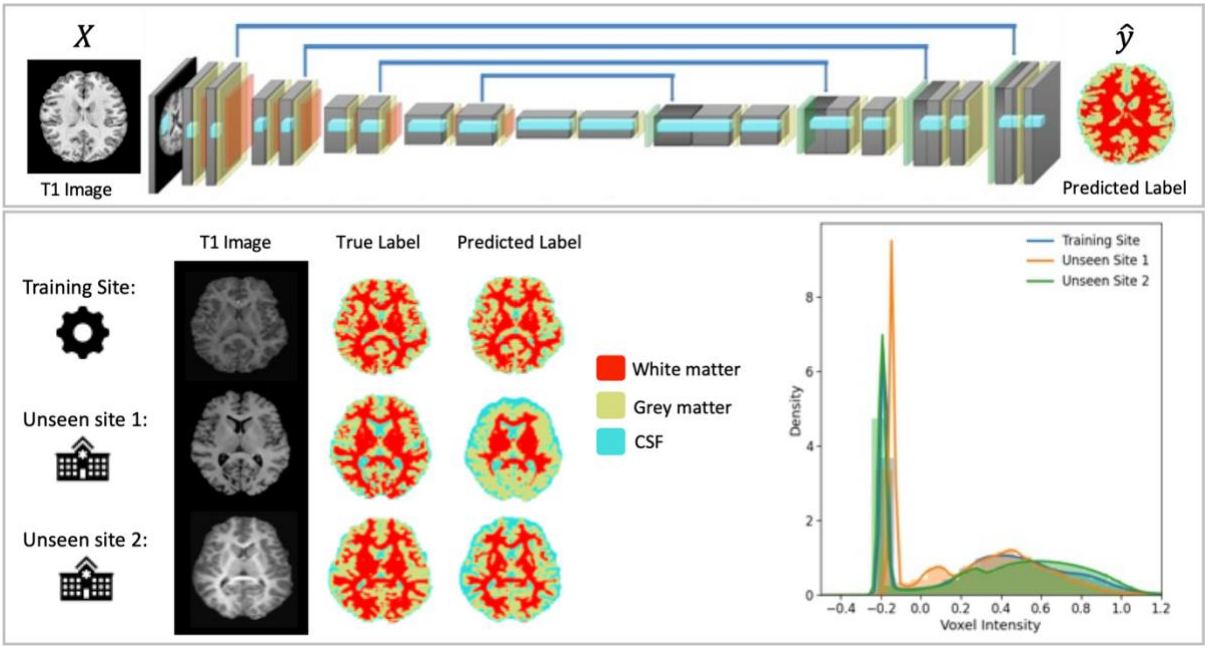

Figure 3: To demonstrate the effect of the difference between domain datasets or domain shift, tissue segmentation was carried out on data from three sites collected as part of the ABIDE (Di Martino, et al. 2013) multisite dataset. Although the data was all collected as part of one study, differences exist between the data collected at different sites due to scanner differences. The architecture used was a 3D UNet (Cicek, et al. 2016) with T1 as the input image; only images from one site were used during training. The predictions can be seen for example images from three sites - one seen during training and two unseen. The segmentation for the site seen during training is good but suffers significant degradation when applied to the unseen sites, despite them being collected for the same study and having similar (normalised) voxel intensities, demonstrating the potential difficulties caused by domain shift.

Having sufficient data to train the model, however, is only the first difficulty for clinical application. The flexibility that allows deep learning methods to learn complex and highly non-linear mappings between the input images and the labels comes at a cost: deep learning methods are prone to *overfitting* to the training data (Srivastava, et al. 2014); this is exacerbated if the amount of training data is insufficient. Further, while a well-trained model should interpolate well to data which falls within the same distribution as that seen during training, the performance degrades quickly once it must extrapolate to out-of-distribution data. Even perturbations unnoticeable to the human eye can cause network performance to collapse (Papernot, et al. 2017). For clinical translatability, we need generalisability from the training set to all other reasonable datasets, including future datasets as yet

uncollected, otherwise a result may be a function of domain drift rather than of the subject's pathology.

Multisite datasets, such as the ABIDE study (Di Martino, et al. 2013), still show an increase in non-biological variance when we pool data across sites and scanners (Yu, et al. 2018). A demonstration of this variance leading to performance degradation for a segmentation task is shown in Fig. 3. Multiple studies have confirmed this variation, identifying causes (batch effects) from scanner and acquisition differences, including scanner manufacturer (Han, et al. 2006), scanner upgrade (Han, et al. 2006), scanner drift (Takao, Hayashi and Ohtomo 2011), scanner magnet strength (Han, et al. 2006), and gradient non-linearities (Jovicich, et al. 2006). The removal of scanner-induced variance is therefore vital for neuroimaging studies, especially if models are to be applied to clinical datasets with a small number of subjects for any given site. Most deep learning approaches either use generative methods to output harmonised versions of the input data (Cetin Karayumak, et al. 2019) (Dewey, et al. 2019), or aim to remove the scanner-related information from the features used to produce the predictions, for instance using adversarial learning (Dinsdale, Jenkinson and Namburete 2021). These methods succeed in removing the scanner effects from the predictions, but hold no guarantees for scanners not seen during training. Further, any harmonised output images are hard to validate without *'travelling heads datasets'* (images from the same subjects acquired on the different scanners) (Moyer, et al. 2020).

The domain shift experienced with multisite data is less than might be expected when we move between research and clinical data, or even just two datasets collected independently. The domain shift here can come from two sources: the scanner and acquisition, and the demographics of the studies. First, MRI scans collected for research are often at a higher resolution and field-strength than clinical scans: clinical scans are designed to be more time efficient, both in terms of the time required for acquisition and for visual inspection, and are often collected at lower resolutions and field strengths. Also, research scans frequently have isotropic voxel sizes, whereas anisotropic voxels are still the norm in the clinic and present in the majority of legacy data (Iglesias, et al. 2020). Unfortunately, due to the aforementioned paucity of training data, we are unlikely to be able to train sophisticated models directly and solely on clinical data in the near future.

Thus, methods being developed that consider this domain shift (e.g., between clinical and research data), focus either on domain adaptation approaches to create shared feature representations for the different datasets, or on synthesising data to enable us to use the clinical domain. Domain adaptation techniques normally consider the situation where there is a large source dataset – e.g. a research dataset such as UK Biobank (Sudlow, et al. 2015) – and a much smaller target dataset – e.g. the clinical dataset of interest, and generally aim to force the learned features to have the same distribution from across sites such that information can be shared across datasets. Domain adaptation approaches have been applied for segmentation (Kamnitsas, Baumgartner, et al. 2017) (Sundaresan, et al. 2021) and classification problems (Guan, et al. 2020). These methods can perform well on the target clinical data, through harnessing information from a large dataset to improve our understanding of a clinical dataset of interest, but further work is required to enable them to adapt reliably to higher numbers of datasets simultaneously.

Domain adaptation methods, at the extreme, essentially have the end goal that the network would work regardless of the acquisition, which is an active area of research (Billot, Greve, et al. 2020) (Thakur, et al. 2020). The other approach which has been explored is to use generative methods to convert the data from one domain to the other (Iglesias, et al. 2020), such that the transformed data can be used in the existing model. Any generated images must be carefully validated to ensure that they convey the same information as the originals and that the outcomes are the same.

3.5 Data Composition and Algorithmic Biases

Finally, we must consider that the demographics of study data frequently do not fully represent the population as a whole, so a domain shift is experienced when we attempt to move from

(e.g.) the research domain to the clinical domain. Because research data is usually acquired with targeted exploration of a certain study question in mind, the datasets rarely contain subjects with co-morbidities or incidental findings. For example, patients with advanced Alzheimer's disease are unlikely to be recruited for a general imaging study, due to ethical implications such as the inability to consent (Clement, et al. 2019). Also, a strong selection bias exists in both recruitment and completion, with studies having demonstrated biases in age, education, ancestry, geographic location, and health status (Clement, et al. 2019). Furthermore, people with family connections to a given condition are more likely to volunteer for a study as a healthy control, leading to certain genetic markers being more prevalent in a study dataset than in the population as a whole (Hostage, et al. 2013). Therefore, associations learned when considering research data may not generalise, and care must be taken in extrapolating any model trained on these datasets to clinical populations.

Models therefore suffer from algorithmic bias: that is, the outcomes of the model may potentially systematically be less favourable to, or have lower performance on, individuals within a particular group, where no relevant difference (e.g. pathology) between groups exists to justify such effects (Paulus and Kent 2020). Erroneous or unsuitable outcomes may be produced for groups less likely to be represented in the training data. As networks simply learn the patterns in the data, any bias in the data may be learned and encoded into the models.

Inevitably, when considering complicated questions with extremely heterogeneous populations, the datasets used to train the deep learning methods will be incomplete and insufficient in terms of spanning all possible modes of variability (Ning, et al. 2020). For instance, pathologies will occur against a background of normal ageing, with differences being present between individuals due to both processes. Sufficiently encompassing all of this variation is infeasible, due both to the number of subjects which would be required, and to the difficulty in recruiting subjects from some specific groups. Thus, when models are developed for clinical translation, the limitations of the models must be understood; wherever groups are under-represented, the appropriateness of the application of the model must be considered, and any limitations identified.

4. Interpretability and Trust

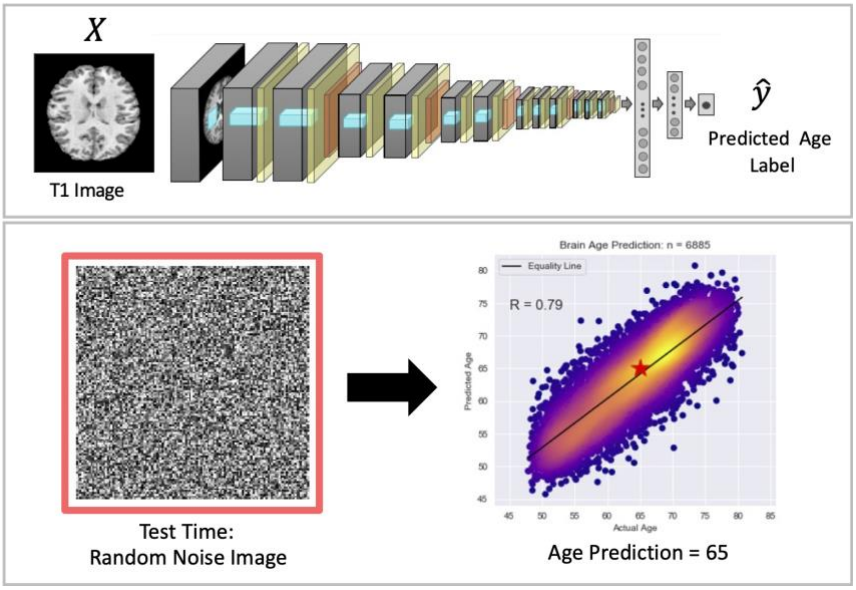

Figure 4: When a model trained to predict brain age (Dinsdale, Bluemke, et al. 2021) from T1 structural images was presented with an image of random noise, as it was unable to output an unknown class, the network predicted the random noise to have an age of 65 - around the average age of the dataset's subjects. While we can easily identify the random noise image by eye, there are many situations where the model, if presented with an image outside of the distribution it was trained on, would still output a (meaningless) result, which would be much harder for a user to identify.

Performance degradation experienced with domain shift would be potentially less problematic were it not for another issue of deep learning methods: models will output a prediction for any data but that prediction may not necessarily be meaningful. Lacking a 'do not know' option, a neural network will output a prediction, even if it is meaningless, or the input nonsensical. For instance, if a random noise image is fed into a network trained to predict brain age, the network will predict an apparently valid age for the random noise (see Fig. 4). Whilst here, visually identifying the pure-noise image is trivial, where the network is trained for a more complicated classification task, identifying erroneous results is more difficult and requires both clinical and domain knowledge, leading to a critical question: can the results be trusted?

The majority would agree that, for deep learning methods to be used to determine patient care, they must be interpretable and interrogable. Interpretability is often defined as *'the ability to provide explanations in understandable terms to a human'*. The explanations should, therefore, be logical decision rules which lead to a given diagnosis or patient care being chosen. This is especially important because neural networks have no semantic understanding of the problem they are being asked to solve. Thus, if spurious information (or *confounders)* in X exists which can aid in this mapping, then this information will probably be used, misleading the predictive potential of the network. Consider for instance, the case where all subjects with a given pathology were collected on the same scanner. A network could then achieve 100% recall accuracy for this pathology by fitting to the scanner signal, rather than learning any information about the pathology (Winkler, et al. 2019). It would then, in all probability, identify a healthy control from the same scanner as having the same pathology.

The effect of confounders would not be observed without further probing the behaviour of the trained model - and probing networks is non-trivial. This has led to neural networks being commonly described as 'blackbox' methods. There is a need for interpretable networks, allowing both understanding and scrutiny of decisions made, which with existing techniques is currently not possible. While this may be acceptable for many computer vision tasks, interrogability is indispensable for clinical neuroimaging tasks. Approaches have been developed to try to enable some insight, which have broadly focused on two main areas: visualisation and uncertainty.

4.1 Visualisation

Visualisation methods generally attempt to show which aspects of the input image led to the given classification – the salient regions - often by creating a `heat map' of importance within the input image. Many of these methods are post-hoc, taking a pre-trained model and testing which regions of the image drove the model prediction. Most commonly, they analyse the gradients or activations of the network for a given input image, such as saliency maps (Simonyan, Vedaldi and Zisserman 2014) or layerwise relevance propagation (Binder, et al. 2016) and have been applied in a range of MRI analysis tasks to explain decision-making.

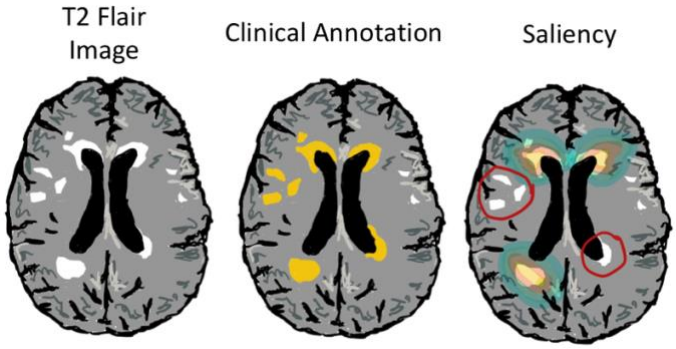

Figure 5: Schematic of the limitation of using saliency: when identifying the presence of white matter hyper-intensities, the neural network might only need to focus on a few of them to make the prediction. Thus, not all of the white matter hyper-intensities would be indicated in the saliency map so the prediction would not match the clinician's expectation.

(Böhle, et al. 2019) and brain age prediction (Dinsdale, Bluemke, et al. 2021). However, concerns remain that these methods do not pass basic sanity checks, and are not providing a valid insight into the model (Adebayo, et al. 2018). Other methods are occlusion- or perturbation-based, where parts of the image are removed or altered in the input, then heat maps are generated which evaluate the effect of this perturbation on the network's performance (Zeiler and Fergus 2014). Most of these methods, however, provide coarse and low-resolution attribution maps and are computationally very expensive (Bass, et al. 2020), especially when working with 3D medical images.

These post-hoc methods do not require any model training in addition to the original network; however, it appears they often fail to identify all the salient regions of a class, especially in medical imaging applications (Bass, et al. 2020). Classifiers base their results on certain salient regions, rather than the object as a whole, and a classifier may therefore ignore a region if the information there is redundant - i.e., if it can be provided by a different region of the image which is sufficient to minimise the loss function. Therefore, the regions of interest highlighted by these methods may not fully match a clinician's expectations (see Fig. 5): also, the prediction results might be virtually unchanged if the network were retrained with supposedly salient areas removed. Generally, although many methods have been developed to produce saliency or 'heatmaps' from CNNs, limited effort has been focused on their evaluation with end-users (Alqaraawi, et al. 2020). Fundamentally, these methods at best only highlight the important content of the image, rather than uncovering the internal mechanisms of the model, and thus only indicate *what* is important, not *why*. Further, they are limited by the fact that CNNs are highly nonlinear systems, so it is unlikely that, in general, there will be a mapping between regions of the input image and the task output that are understandable to humans.

Attention gates are components of the network which aim to focus a CNN on the target region of the image (the salient regions) by suppressing irrelevant feature responses in feature maps *during* training rather than post-hoc (Park, et al. 2018). This provides the user with attention maps, which again highlight the regions of the input image driving the network predictions. However, these methods, similarly to saliency or gradient-based methods, may not highlight all the expected regions in the image, and can only indicate regions, not elucidate why. Attention gates have been applied to a range of imaging tasks, both for classification (Dinsdale, Bluemke, et al. 2021) and segmentation (Schlemper, et al. 2019). Other methods have been developed to allow the visualisation of the differences between classes directly, rather than analysing the model post-hoc (Bass, et al. 2020) (Lee, et al. 2020).

The methods discussed so far enable visualisation of the regions of the input image that drive the predictions, but do not provide insight into *how* the underlying filters of the network create decision boundaries, or why the regions were important, and are vulnerable to confirmation bias. In addition, in neuroimaging, patients with a given pathology are typically heterogeneous, and any changes they cause probably occur simultaneously. There are also significant amounts of healthy and normal variation in shape and appearance, so the interpretation of feature attribution maps to understand network predictions is difficult. Given the millions of parameters in many deep learning networks, despite our ability to visualise individual filters and weights helping us to understand the hierarchical image composition, it is difficult to interrogate why decisions were made. Without some understanding of the model decision-making process, application across neuroimaging tasks in a clinical setting is likely to be limited, due to the lack of trust that could be placed on the decisions. This is less of a concern in some settings, such as lesion segmentation, where the outputs can potentially be validated manually, but for tasks such as disease prediction, there may be greater concerns about model interpretability, which are yet to be solved by existing approaches.

4.2 Uncertainty

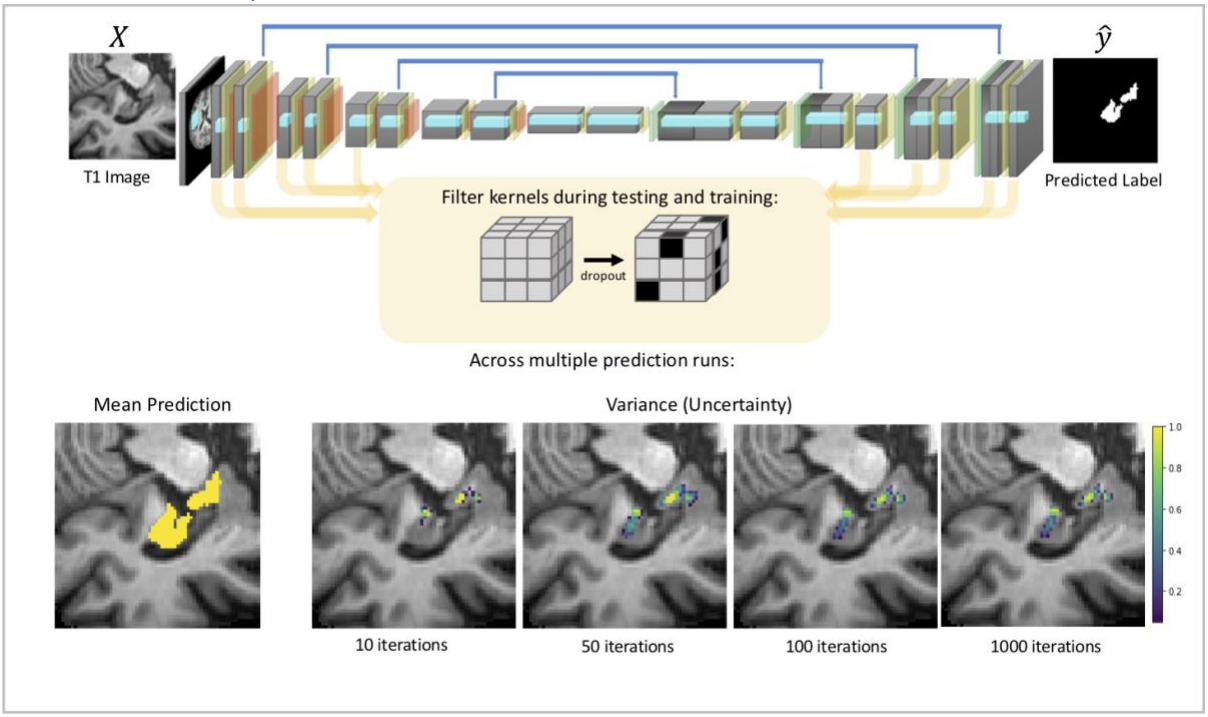

Figure 2: Most uncertainty methods have dropout applied at training and test time. Weights in the convolutional kernels are removed, which is approximated to represent the distribution of possible model architectures at test time. To demonstrate this, we trained a standard 3D UNet to complete hippocampal segmentation, with a dropout value of 0.5 applied on all convolutional layers. The HarP dataset (Frisoni and Jack 2015) was used in this experiment, pre-processed as in (Dinsdale, Jenkinson and Namburete 2019). For each subject, we obtained a mean prediction and an uncertainty map, indicating the regions where the predictions between models were most varied and so approximated to be least certain.

The use of uncertainties is an approach which aims to address the problem that, regardless of the input image, neural networks will always output a prediction, however inaccurate. Thus, by providing an estimate of the uncertainty associated with the prediction we can help the user to make an informed decision about whether or not to trust the model prediction. The softmax values output by a neural network are not true probabilities (Gal and Ghahramani 2016), and networks often output high, incorrect softmax values, especially when presented with noisy or ambiguous data, or when the data presented to them differs from the distribution of the training data, so uncertainties are needed to allow proper quantification of the confidence of the prediction.

Uncertainties in deep learning can be split into two distinct groups (Kendall 2017): *aleatoric uncertainty*, due to the ambiguity and noise in the data, and *epistemic uncertainty*, due to the uncertainty in the model parameters. The majority of methods in the literature focus on epistemic uncertainty, using Bayesian approaches to quantify the degree of uncertainty. The goal here is to estimate the posterior distribution of the model parameters. However, due to the very high dimensional parameter space, analytically computing the posterior directly is infeasible. Therefore, most methods use Monte Carlo dropout (Gal and Ghahramani 2016), where dropout is applied to each of the convolutional layers and kept at test time; thus, we are able to sample from the distribution of possible model architectures. The uncertainty is then quantified through the variance of the predictive distribution, resulting from multiple iterations of the prediction stage with dropout present at test time, as demonstrated in Fig. 6. This approach can readily be applied to existing convolutional neural networks; in medical imaging it has primarily been used for segmentation tasks (Roy, Navab and Wachinger 2018), where the segmentation is predicted alongside an uncertainty map. Other works have studied disease prediction, where the uncertainty is associated with the predicted class (Tousignant, et al. 2019), and image registration (Bian, et al. 2020). However, care must be taken with choice of the hyperparameters to ensure that the model assumptions are reasonable.

Some methods focus on the aleatoric uncertainty instead, estimated by having augmentation at test time (Ayhan and Berens 2018) (Wang, et al. 2019). Understanding of the uncertainty introduced by data varying from the training distribution is vital for clinical translation of deep learning techniques. Given the degree of variation present in clinical data between sites and scanners, it is vital to understand how this contributes to predictions, both to mitigate against it, and to develop user confidence in the predictions. Correlation between erroneous predictions and high uncertainties exists, so this could be used to improve the eventual predictions (Jungo, et al. 2018).

However, further work in this area is still needed to ensure that the uncertainties produced would be meaningful at deployment, for instance across dataset shifts. Calibration of uncertainties is also necessary so that they are comparable across methods (Thagaard, et al. 2020). Furthermore, uncertainty values are only as good as the model and only meaningful alongside a well-validated model which is sufficiently powerful to discriminate the class of interest.

4.3 Interrogating the Decision Boundary

For many applications in neuroimaging, the output of a deep learning algorithm, if applied clinically, could potentially directly influence patient care and outcomes. Thus, there is a clear need to be able to interrogate how decisions were made (Shah, Milstein and Bagley 2019). While visualisation methods allow inspection of which regions of the image influenced the prediction, and uncertainties grant us insight as to the confidence we should place in a prediction, for many applications we need to know precisely which characteristics led to a given prediction and what would need to change for the outcome to be different, and to help identify any bias driving predictions.

Our ability to interrogate the decision boundary is currently limited. *Counterfactual analysis* is one of the few existing approaches, which, given a supervised model where the desired prediction has not been achieved, shows what would have happened if the input were altered slightly (Verma, Dickerson and Hines 2020). Simply, it identifies what altered characteristics would have led to a different model prediction. However, applications to neuroimaging (Pawlowski, Coelho de Castro and Glocker 2020) are currently few and exploration of its utility across neuroimaging tasks is required to ascertain its viability in a clinical setting.

5. Evaluation

5.1 Availability of Training Labels

The evaluation of metrics requires labels: the *ground truth*. We generally regard the ground truth as labels created by *domain experts*; these labels are key for training models, but do not necessarily form part of standard clinical practice. Labels are required both for evaluation of the model performance and to train supervised methods. This exacerbates the problem of the shortage of data as we need both large amounts of data, and equal amounts of labels. These labels are expensive to obtain, requiring large allocation of expert time to curate and expert domain knowledge, and are unlikely to be available for every clinical imaging site. Thus, we need methods which work when low numbers of labelled data points are available.

Few- and zero-shot learning methods work in very low-data regimes and are beginning to be applied to medical imaging problems (Feyjie, et al. 2020). They are unlikely to generalise well to images from other sites and scanners, as the variation seen will not span the expected variation of the data, but they can help to begin to learn clusters of similar subjects where few labels are available. Unsupervised domain adaptation has been applied more widely, including for neuroimaging problems, to help to cope with a lack of labels, with information from one dataset being leveraged to help us perform the same or a related task on another dataset (Sundaresan, et al. 2021).

Other methods to overcome the lack of available labels focus on working with approximations for labels, which are cheaper to acquire (Tajbakhsh, et al. 2020). Many methods propose pre-training the network using auxiliary labels generated using automatic tools and then fine-tuning the model on the small number of manual labels (Guha Roy, et al. 2018), or registration of an atlas to propagate

labels from the atlas to the subject space (Hesse, et al. 2022). Other approaches are weakly supervised, utilising quick annotations such as image-level labels (Feng, et al. 2017) or bounding box annotations.

Other approaches to allow us to utilise deep learning when we have limited numbers of training labels include active learning and omnisupervised learning, both trying to make the most effective use of the limited number of labels available. Active learning aims to minimise the quantity of labelled data required to train the network by prompting a human labeller to produce additional manual labels only where they might provide the greatest performance improvements, thereby minimising the total number of annotations that need to be provided, but giving better performance than random annotation of the same number of samples (Yang, et al. 2017). In omnisupervised learning (Radosavovic, et al. 2018) automatically generated labels are created to improve predictions, starting from a small, labelled training set. By combining data diversity through applying data augmentation, and model diversity through the use of multiple different models, a consensus of labels is produced, which can be used to train the final model (Huang, Noble and Namburete 2018).

The difficulty in acquiring good quality manual labels is exacerbated by the variance caused when we pool data. The labels themselves provide an additional source of variance: when working in neuroimaging, the labels are frequently complicated and ambiguous (Shwartzman, et al. 2020), often open to interpretation or with subjects having multiple labels that could be attributed due to co-morbidities (Graber 2013). Despite this, we usually assume them to be 100% accurate (Cabitza, et al. 2020) - the 'gold' standard. If there is no objective answer, we cannot expect networks to provide one. Furthermore, this also leads to *inter-rater* variability, which generates a degree of uncertainty in the produced ground truth. The effect that this variability has on the predictions of the network needs to be understood and mitigated against. The uncertainty in the labels is also amplified by the lack of available data for rare conditions, which are therefore less represented in datasets, resulting in raters having less experience assessing them - particularly problematic if trying to quantify longitudinal changes with different raters (Visser, et al. 2019).

Approaches need to consider three factors (Cabitza, et al. 2020): *agreement* -- the degree to which raters agree on a given label; *confidence* -- how certain a rater is in their label, and *competence* - how accurate a given rater is. Research directions into the effect of rater variance have largely focused either on quantifying the reliability of the labels (Cabitza, et al. 2020), or quantifying its effect on network performance (Shwartzman, et al. 2020). Before any algorithm is deployed in practice, the limitations due to the labels must be understood, and its consideration become a standard part of any deployment pipeline, remembering the *'garbage in, garbage out'* principle.

5.2 Choice of Loss Function

When training and evaluating model performance, we must choose a loss, or cost, function which we aim to minimise. Although some works design bespoke, task-specific cost functions, the majority are based on standard functions, such as categorical cross entropy for classification and segmentation tasks, Dice (an overlap metric) for segmentation and mean square error (MSE) for regression-based tasks.

These metrics are normally chosen because of their well-understood and characterised behaviour (Maier-Hein, et al. 2018). For clinical translation of deep learning methods, we need to consider which measures are most important for the clinical application (Shah, Milstein and Bagley 2019) (Keane and Topol 2018). Metrics only tell us part of the story: it is crucial to ensure that all vital information for clinical assessment is provided by the reported metrics. For instance, in many cases, false negatives are more problematic than false positives, resulting in a patient failing to receive the necessary care. Developing networks and loss functions with each specific application in mind is vital.

Furthermore, when training neural networks, we generally maximise the average performance. In practice, however, we are more likely to care about the performance on the hardest examples being acceptable, than the performance on the easiest set of examples being improved slightly (Shu, et al. 2020). Trading-off a small amount of performance on easier examples in return for better performance on harder examples, which may give the same average performance overall, is

probably preferable. Thus, the standard practice of minimising the average performance may not be appropriate.

6. Logistical Challenges

6.1 Computational Resource

The final category of challenges is more logistical. Many of the most successful methods applied in imaging challenges involve large ensemble models such as the nnU-Net (Isensee, et al. 2021), leading to many parameters and, therefore, calculations that must be stored and computed. While successful in challenges, these methods are often not implementable on the hardware available in practice. Therefore, for clinical translatability, methods need to be developed which consider that computational limitations will be present on deployment and seek to create solutions which work within these constraints. Student-teacher networks (Hinton, Vinyals and Dean 2015) and model distillation (Murugesan, et al. 2020) aim to create smaller networks capable of mimicking the performance of the original large model (teacher), thus reducing the number of parameters in the final network which is deployed (student). Other approaches use separable convolutions which drastically reduce the number of parameters in the network. Model pruning (LeCun, Denker and Solla, 1990) (Dinsdale, Jenkinson and Namburete 2021b) acknowledges that the parameters in neural networks are sparse and, therefore, by removing those that contribute least to the final prediction, we can reduce the size of the model architecture whilst maintaining performance.

6.2 Data Sharing and Data Privacy

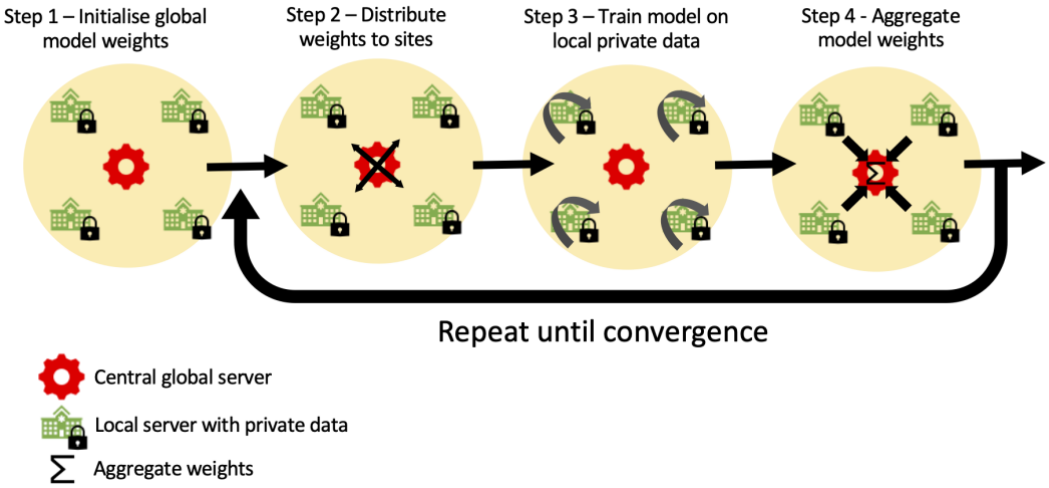

Figure 7: Illustration of a centralised federated learning framework. In the framework, the data for training the model is stored in local servers and not shared with the central server to ensure data security. While the global model is available in the central server, the model parameters are shared with the local nodes 1,2... N where training and parameter updates happen. The updated weights are then received at the central server, where the incoming updates are aggregated and applied to the global model. This learning and update happens in an iterative manner; both up and down transfer of model parameters are encrypted for data security.

If we want CNNs that work for patients in real clinical applications, we need to be able to train our models on medical data that are relevant, realistic, and representative. Many current approaches focus on pooling anonymised data from across sites and patient groups through removing identifiable features such as name, birth date and faces from the images. However, neural networks are still capable of extracting identifiable features from these anonymised images such as age and sex

(Pawlowski, Coelho de Castro and Glocker 2020), which, in combination with other features such as hospital location and illness, could be identifying (Sweeney 2002). The ability of neural networks to extract this information is only likely to increase. Furthermore, a proportion of identification risk

comes from the presence of other auxiliary information - for instance, in neuroimaging, the scanner used to acquire the image. This is known as *linkage attack* and is increasingly difficult to protect against fields using classic anonymisation techniques (Sweeney 2002).

While de-identifying these data may just seem like an extra task for medical researchers, there are parties whose core business model is to de-anonymize medical data that have been sold for research purposes and sell that information to insurance companies (Tanner 2017). De-anonymisation research is a rapidly advancing field - for instance, reconstructing the faces of defaced medical images (Abramian and Eklund 2019). Thus, to avoid future data privacy problems, approaches avoiding the aggregation of private medical information are valuable.

Fortunately, medical research is not the only field to face difficulties regarding the handling of sensitive, personal information. For instance, banking and mobile phone companies have faced this problem before. Therefore, we can take advantage of the privacy-preserving data analysis techniques that have rapidly developed in recent years. These techniques allow models to be trained without having direct access to the data, and prevent these models from inadvertently storing sensitive information about the data. The most popular of these techniques are *federated learning, differential privacy*, and various forms of encrypted computation (Al-Rubaie and Chang 2019) (Kaissis, et al. 2021). Here we will focus on federated learning and differential privacy, as they currently show the most practical relevance in a neuroscience research setting (Rieke, et al. 2020).

Federated learning (Fig. 7) means training or testing your model on data that is stored on different devices or servers across the world, without having to centrally collect the data samples into one local aggregate dataset (Li, et al. 2020). Instead of moving the data to the model, copies of the global model are sent to where the data is located; the data remains on the hospital server. The model is then trained on the local data, after which the newly improved model with its updated parameters is sent back to the main server to be aggregated with the main model. This preserves privacy in the sense that the data has not been moved from the device, and is therefore gaining popularity in various healthcare applications (Sheller, et al. 2019). However, federated learning is limited by the fact that the content of the local data can sometimes be inferred from the weight updates or improvements in the models (Wang, et al. 2019) or due to the large numbers of parameters memorising information about individuals.

Differential privacy helps overcome these drawbacks by injecting statistical noise to obscure the data contributions from individuals in the dataset (Dwork and Roth 2014) (Ziller, et al. 2021). This is performed while ensuring that the model still gains insight into the overall population, and thus provides predictions that are accurate enough to be useful. Ultimately, the use of differential privacy is a careful trade-off between privacy preservation and model utility (Dwork and Roth 2014). A critical aspect of differential privacy is its inherent robustness to linkage attacks (Sweeney 2002). As methods are developed, consideration of these approaches, and future developments will be vital for ensuring privacy is maintained.

7. Conclusion

The combination of deep learning-based methods and large-scale imaging datasets, such as UK Biobank, offers many opportunities to neuroimaging. Clearly, however, for the full impact of these methods to be experienced in the clinical domain there are challenges that must still be overcome. Our key recommendations for future directions are discussed in Box 1. Ultimately, for models to be able to be deployed successfully, the clinical needs and limitations must be considered central to model design, so that the models produced are robust, reliable, and able to improve patient outcomes. In this article, we have discussed issues relating to data availability, interpretability, model evaluation and data privacy. Deep learning-based methods are beginning to receive FDA approval for applications in medical imaging, but it is yet to be seen what impact or uptake these methods will have. The challenges for neuroimaging are, however, likely to differ in focus to those of the computer vision field. In particular, interpretability - the ability to interrogate decision making, and trust the

decision-making process - is likely to be a significant barrier for translatability and will likely require specific efforts beyond those in the general computer vision field.

The code for the examples in this paper can be found at: github.com/nkdinsdale/challenges_review.

Box 1: Recommendations for Future Directions

Throughout this review, we have discussed the key barriers for the success of deep learning in neuroimaging, and the current approaches and directions being explored to overcome them. Although various methods are being explored, the barriers remain significant and thus, we here briefly discuss our recommendations for future research directions.

- **Data Availability**: Current challenges are interlinked to the data available: inevitably, the data we have collected can only ever be a snapshot of the populations we wish to study. We need to better understand the limitations of the data we have available, for instance, through risk analysis (Zendle, et al. 2015) to identify the underrepresented demographics in the data. Where underrepresented groups or other forms of training-sample-bias are identified, this should enable exploration of mitigation approaches, such as oversampling underrepresented groups, creating simulated subjects using generative methods or targeted data collection.
- **Data pooling and harmonization**: The relative cost and difficulty of acquiring imaging data (compared with simpler, smaller forms of subject-level data such as simple demographics) makes it hard to build up large-N imaging datasets for training deep-learning models; furthermore, the cost, size and complexity of imaging data further exacerbates this. It will often be necessary therefore to pool datasets, creating privacy and harmonisation issues. More work on robust multi-modal image processing, applied before deep-learning training, will be needed to reduce problems of data harmonization (for example, reducing variations due to imaging hardware and acquisition protocol). However, as this is unlikely to be perfect, deep-learning models will still likely need to include a harmonization component, an important area of future research.
- **Interpretability and Trust:** We need to develop better methods which explain *why* a prediction has been made -- in addition to *what* drives the predictions. Until it is possible to train truly interpretable deep learning networks, in safety-critical applications, such as methods to predict diagnosis or suggest treatment options, methods should be used which are inherently interpretable, such that patients and clinicians can interrogate outcomes, or we risk long-term damage to the trust in deep learning models (Rudin 2019).
- **Evaluation:** After training, the model evaluation must be thorough, to ensure that the model works as expected and is robust to the expected variation. Better robust evaluation procedures should be developed which maximize the impact of the available labels, for instance through test time augmentation to simulate variation (Hendryck and Dietterich 2019), multiple evaluation metrics to capture different aspects of the prediction, or stratifying results to understand performance across different demographic groups such as age or sex.
- **Logistical Challenges:** The use of representative clinical data will be vital for the success of deep learning models; thus, privacy-preserving approaches are important for new methodological development. Therefore, future developments should be built around federated frameworks (i.e., non-centralized data stores), despite the increased constraint on architectures and training procedures (Dinsdale, Jenkinson and Namburete 2022). New methods will need to minimize the amount of information which needs to be shared (to the centralized training process), and properly understand the dangers of de-anonymization (e.g., understanding differential privacy).

Acknowledgements

This work was supported in parts by funding from the Engineering and Physical Sciences Research Council (EPSRC) and Medical Research Council (MRC) [grant number EP/L016052/1] (ND and EB), by the Clarendon Scholarship fund (EB), and a Wellcome Trust Collaborative Award (215573/Z/19/Z) (SS). The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust (203139/Z/16/Z) (VS). A.N. is grateful for support from the UK Royal Academy of Engineering under the Engineering for Development Research Fellowships scheme, ND and AN are also supported

by an Academy of Medical Sciences Springboard Grant (SBF005/1136). M.J. is supported by the National Institute for Health Research (NIHR) and the Oxford Biomedical Research Centre (BRC).

This research has been conducted in part using the UK Biobank Resource under Application Number 8107. We are grateful to UK Biobank for making the data available, and to all UK Biobank study participants, who generously donated their time to make this resource possible. Analysis was carried out on the clusters at the Oxford Biomedical Research Computing (BMRC) facility. BMRC is a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute, supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Declaration of Interests

Mark Jenkinson receives royalties from licensing of FSL to non-academic, commercial parties.

Bibliography

Abadi, Martin, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. "Deep Learning with Differential Privacy." *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* 308–318.

Abramian, David, and Anders Eklund. 2019. "2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)." *Refacing: Reconstructing Anonymized Facial Features Using GANS* 1104-1108.

Abrol, Anees, Manish Bhattarai, Alex Fedorov, Yuhui Du, Sergey Plis, and Vince Calhoun. 2020. "Deep residual learning for neuroimaging: An application to predict progression to Alzheimer's disease." *Journal of Neuroscience Methods* 339: 108701.

Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. "Sanity Checks for Saliency Maps." *32nd Conference on Neural Information Processing Systems (NeurIPS 2018).*

Alqaraawi, Ahmed, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Bianchi-Berthouze. 2020. "Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study." *IUI '20: 25th International Conference on Intelligent User Interfaces* 275-285.

Al-Rubaie, Mohammad, and J. Morris Chang. 2019. "Privacy-Preserving Machine Learning: Threats and Solutions." *IEEE Security Privacy* 49-58.

Ayhan, M, and Philipp Berens. 2018. "Test-time Data Augmentation for Estimation of Heteroscedastic Aleatoric Uncertainty in Deep Neural Networks." *Proceedings of Medical Imaging with Deep Learning (2018).*

Böhle, Moritz, Fabian Eitel, Martin Weygandt, and Kerstin Ritter. 2019. "Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification." *Frontiers in Aging Neuroscience* 194.

Bashyam, Vishnu, Guray Erus, Jimit Doshi, Mohamad Habes, Ilya Nasralah, Monica Truelove-Hill, Dhivya Srinivasan, et al. 2020. "MRI signatures of brain age and disease over the

lifespan based on a deep brain network and 14 468 individuals worldwide." *Brain : a journal of neurology* 143.

Bass, Cher, MD Silva, CH Sudre, PD Tudosiu, Steve M Smith, and Emma C Robinson. 2020. "ICAM: Interpretable Classification via Disentangled Representations and Feature Attribution Mapping." *34th Conference on Neural Information Processing Systems (NeurIPS 2020).*

Baumgartner, C. F., L. M. Koch, K. C. Tegzan, J. X. Ang, and E Konukoglu. 2018. "Visual Feature Attribution Using Wasserstein GANs." *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 8309-8319.

Baumgartner, Christian F., Kerem C. Tezcan, Krishna Chaitanya, Andreas M. Hötker, Urs J. Muehlematter, Khoschy Schawkat, Anton S. Becker, Olivio Donati, and Ender Konukoglu. 2019. "PHiSeg: Capturing Uncertainty in Medical Image Segmentation." *Medical Image Computing and Computer Assisted Intervention -- MICCAI 2019* 119--127".

Ben-David, Shai, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. 2010. "A theory of learning from different domains." *Machine Learning* 151--175.

Bian, Cheng, Chenglang Yuan, Jiexiang Wang, Meng Li, Xin Yang, Shuang Yu, Kai Ma, Jin Yuan, and Yefeng Zheng. 2020. "Uncertainty-aware domain alignment for anatomical structure segmentation." *Medical Image Analysis* 101732.

Billot, Benjamin, Douglas Greve, Koen Van Leemput, Bruce Fischl, Juan Iglesias, and Adrian Dalca. 2020. "A Learning Strategy for Contrast-agnostic MRI Segmentation." *Proceedings of Machine Imaging with Deep Learning (MIDL) 2020.*

Billot, Benjamin, Eleanor Robinson, Adrian V Dalca, and Juan Eugenio Iglesias. 2020. "Partial Volume Segmentation of Brain {MRI} Scans of Any Resolution and Contrast." *Medical Image Computing and Computer Assisted Intervention -- MICCAI 2020.*

Billot, Benjamin, Martina Bocchetta, Emily Todd, Adrian V. Dalca, Jonathan D. Rohrer, and Juan Eugenio Iglesias. 2020. "Automated segmentation of the hypothalamus and associated subunits in brain MRI." *NeuroImage* 117287.

Binder, Alexander, Sebastian Lapuschkin, Gregoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. "Layer-Wise Relevance Propagation for Deep Neural Network Architectures." *Information Science and Applications (ICISA) 2016* 913-922.

Bindschaedler, Vincent, Paul Grubbs, David Cash, Thomas Ristenpart, and Vitaly Shmatikov. 2018. "The Tao of Inference in Privacy-Protected Databases." *Proc. VLDB Endow.* 1715–1728.

Bookheimer, Susan, David Salat, Melissa Terpstra, Beau Ances, Deanna Barch, Randy Buckner, Gregory Burgess, et al. 2018. "The Lifespan Human Connectome Project in Aging: An overview." *NeuroImage* 185.

Cabitza, Federico, Andrea Campagner, Domenico Albano, Alberto Aliprandi, Alberto Bruno, Vito Chianca, Angelo Corazza, et al. 2020. "The Elephant in the Machine: Proposing a New Metric of Data Reliability and its Application to a Medical Case to Assess Classification Reliability." *Applied Sciences.*

Cetin Karayumak, Suheyla, Sylvain Bouix, Lipeng Ning, Anthony James, Tim Crow, Martha Shenton, Marek Kubicki, and Yogesh Rathi. 2019. "Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters." *NeuroImage* 184: 180-200.

Chaitanya, Krishna, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. 2020. "Contrastive learning of global and local features for medical image segmentation with limited annotations." *ArXiv.*

Chaitanya, Krishna, Neerav Karani, Christian F. Baumgartner, Anton Becker, Olivio Donati, and Ender Konukoglu. 2019. "Semi-supervised and Task-Driven Data Augmentation." *Information Processing in Medical Imaging* 29-41.

Chapelle, Olivier, Jason Weston, L\'{e}on Bottou, and Vladimir Vapnik. 2001. "Vicinal Risk Minimization." *Advances in Neural Information Processing Systems* 13.

Chen, Chaofan, Oscar Li, Chaofan Tao, Alina Barnett, Jonathan Su, and Cynthia Rudin. 2019b. "This Looks Like That: Deep Learning for Interpretable Image Recognition." *NeurIPS.*

Chen, Hao, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng. 2018. "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images." *NeuroImage* 446-455.

Chen, Liang, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. 2019. "Self-supervised learning for medical image analysis using image context restoration." *Medical Image Analysis* 58: 1361-8415.

Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. "A Simple Framework for Contrastive Learning of Visual Representations." *ArXiv.*

Cho, Junghwan, Kyewook Lee, Ellie Shin, G. Choy, and Do Synho. 2015. "How much data is needed to train a medical image deep learning system to achieve necessary high accuracy." *arXiv.*

Cicek, Ozgun, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 2016. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation." *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016* 424 - 432.

Clement, Clare, Lucy Selman, Patrick Kehoe, Beth Howden, Athene Lane, and Jeremy Horwood. 2019. "Challenges to and Facilitators of Recruitment to an Alzheimer's Disease Clinical Trial: A Qualitative Interview Study." *Journal of Alzheimer's Disease* 1-9.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. "ImageNet: A large-scale hierarchical image database." *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248-255.

Dewey, Blake, Can Zhao, Jacob Reinhold, Aaron Carass, Kathryn Fitzgerald, Elias Sotirchos, Shiv Saidha, et al. 2019. "DeepHarmony: A deep learning approach to contrast harmonization across scanner changes." *Magnetic Resonance Imaging* 64.

Di Martino, Adriana, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco Castellanos, Kaat Alaerts, Jeffrey Anderson, et al. 2013. "The Autism Brain Imaging Data Exchange: Towards Large-Scale Evaluation of the Intrinsic Brain Architecture in Autism." *Molecular psychiatry.*

Dinsdale, Nicola K, Emma Bluemke, Stephen Smith, Zobair Arya, Diego Vidaurre, Mark Jenkinson, and Ana IL Namburete. 2021. "Learning patterns of the ageing brain in MRI using deep convolutional network." *Neuroimage* 224.

Dinsdale, Nicola K, Mark Jenkinson, and Ana IL Namburete. 2022. "FedHarmony: Unlearning Scanner Bias with Distributed Data." *ArXiv.*

Dinsdale, Nicola K., Dinsdale Jenkinson, and Ana I.L. Namburete. 2021. "Deep learning-based unlearning of dataset bias for {MRI} harmonisation and confound removal." *NeuroImage* 117689.

Dinsdale, Nicola K., Mark Jenkinson, and Ana I. L. Namburete. 2019. "Spatial Warping Network for 3D Segmentation of the Hippocampus in MR Images." *Medical Image Computing and Computer Assisted Intervention -- MICCAI 2019* 284--291.

Dinsdale, Nicola K., Mark Jenkinson, and Ana I. L. Namburete. 2021b. "STAMP: Simultaneous Training and Model Pruning for Low Data Regimes in Medical Image Segmentation." *BioArXiv.*

Dorent, Reuben, Samuel Joutard, Jonathan Shapey, Sotirios Bisdas, Neil Kitchen, Robert Bradford, Shakeel Saeed, Marc Modat, Sebastien Ourselin, and Tom Vercauteren. 2020. "Scribble-Based Domain Adaptation via Co-segmentation." *Medical Image Computing and Computer Assisted Intervention -- MICCAI 2020* 479--489.

Dwork, Cynthia, and Aaron Roth. 2014. "The Algorithmic Foundations of Differential Privacy." *Found. Trends Theor. Comput. Sci.* 211-407.

Feng, Xinyang, Jie Yang, Andrew Laine, and Els Angelini. 2017. "Discriminative Localization in CNNs for Weakly-Supervised Segmentation of Pulmonary Nodules." *ArXiv.*

Feyjie, Abdur R, Reza Azad, Marco Pedersoli, Claude Kauffman, Ismail Ben Ayed, and Jose Dolz. 2020. "Semi-supervised few-shot learning for medical image segmentation." *ArXiv.*

Frazier, Jean, Steven Hodge, Janis Breeze, Anthony Giuliano, Janine Terry, Constance Moore, David Kennedy, et al. 2008. "Diagnostic and Sex Effects on Limbic Volumes in Early-Onset Bipolar Disorder and Schizophrenia." *Schizophrenia bulletin* 34: 37-46.

Frisoni, Giovanni, and Clifford Jack. 2015. "HarP: The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation. A standard of reference from a global working group." *Alzheimer's and Dementia.*

Gal, Yarin, and Zoubin Ghahramani. 2016. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." *Proceedings of The 33rd International Conference on Machine Learning* 1050--1059.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. 2015. "Domain-Adversarial Training of Neural Networks." *Journal of Machine Learning Research.*

Gentili, Michele, Sara Hajian, and Carlos Castillo. 2017. "A Case Study of Anonymization of Medical Surveys." *Proceedings of the 2017 International Conference on Digital Health* 77-81.

Ghafoorian, Mohsen, Jonas Teuwen, Rashindra Manniesing, Frank-Erik Leeuw, Bram Ginneken, Nico Karssemeijer, and Bram Platel. 2018. "Student Beats the Teacher: Deep Neural Networks for Lateral Ventricles Segmentation in Brain MR." *Proceedings Volume 10574, Medical Imaging 2018: Image Processing.*

Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. 2011. "Deep Sparse Rectifier Neural Networks." *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* 15.

Graber, Mark L. 2013. "The incidence of diagnostic error in medicine." *BMJ Quality & Safety* 21-27.

Grill, Jean-Bastien, Florian Strub, Florent Altch\'{e}, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo and Guo, Zhaohan Avila Pires, Mohammad Gheshlaghi Azar, and Bilal Piot. 2020. "Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning." *Advances in Neural Information Processing Systems* 33: 21271--21284.

Guan, Hao, Erkun Yang, Pew-Thian Yap, Dinggang Shen, and Mingxia Liu. 2020. "Attention-Guided Deep Domain Adaptation for Brain Dementia Identification with Multi-site Neuroimaging Data." *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning* 31-40.

Guha Roy, Abhijit, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. 2018. "QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy." *NeuroImage.*

Hadsell, Raia, Sumit Chopra, and Yann Lecun. 2006. "Dimensionality Reduction by Learning an Invariant Mapping." *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* 1735 - 1742.

Han, Xiao, Jorge Jovicich, David Salat, Andre Kouwe, Brian Quinn, Silvester Czanner, Evelina Busa, et al. 2006. "Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer." *NeuroImage* 108-194.

Hatamizadeh, Ali, Demetri Terzopoulos, and Andriy Myronenko. 2019. "End-to-End Boundary Aware Networks for Medical Image Segmentation." *Machine Learning in Medical Imaging, 10th International Workshop, MLMI 2019* 187-194.

He, Kaiming, Ross Girshick, and Piotr Dollar. 2019. "Rethinking ImageNet Pre-Training." *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* 4917-4926.

He, Tong, Ru Kong, Avram J. Holmes, Minh Nguyen, Mert R. Sabuncu, Simon B. Eickhoff, Danilo Bzdok, and Jiashi and Yeo, B.T. Thomas Feng. 2020. "Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics." *NeuroImage* 206: 116276.

Henaff, Olivier. 2020. "Data-Efficient Image Recognition with Contrastive Predictive Coding." *Proceedings of the 37th International Conference on Machine Learning* 119: 4182--4192.

Hendryck, Dan, and Thomas Dietterich. 2019. "Benchmarking Neural Network Robustness to Common Corruptions and Pertubations." *ICLR.*

, Leonie, Sailesh Conjeti, Santiago Estrada, Kersten Diers, Bruce Fischl, and Martin Reuter. 2020. "FastSurfer - A fast and accurate deep learning based neuroimaging pipeline." *NeuroImage* 117012.

Herzog, Lisa, Elvis Murina, Oliver Dürr, Susanne Wegener, and Beate Sick. 2020. "Integrating uncertainty in deep neural networks for {MRI} based stroke analysis." *Medical Image Analysis* 101790.

Hesse, Linde S, and Ana IL Namburete. 2022. "INSightR-Net: Interpretable Neural Network for Regression using Similarity-based Comparisons to Prototypical Examples." *ArXiv.*

Hesse, Linde, Moska Aliasi, Felipe Moser, Monique C. Haak, Weidi Xi, Mark Jenkinson, and Ana I.L. Namburete. 2022. "Subcortical segmentation of the fetal brain in 3D ultrasound using deep learning." *NeuroImage* 254.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. "Distilling the Knowledge in a Neural Network." *ArXiv.*

Holmes, Avram, Marisa Hollinshead, Timothy O'Keefe, Victor Petrov, Gabriele Fariello, Lawrence Wald, Bruce Fischl, et al. 2015. "Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures." *Scientific data* 2.

Hostage, Christopher, Kingshuk Choudhury, Pudugramam Doraiswamy, and Jeffrey Petrella. 2013. "Dissecting the Gene Dose-Effects of the APOE eta4 and eta2 Alleles on Hippocampal Volumes in Aging and Alzheimer's Disease." *PLOS One.*

Huang, Li, Andrew L. Shea, Huining Qian, Aditya Masurkar, Hao Deng, and Dianbo Liu. 2019. "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records." *Journal of Biomedical Informatics* 103291.

Huang, Ruobing, J Alison Noble, and Ana Namburete. 2018. "Omni-supervised learning: scaling up to large unlabelled medical datasets." *International Conference on Medical Image Computing and Computer-Assisted Intervention* 572-580.

Hughes, Emer, Tobias Winchman, Francesco Padormo, Rui Pedro Teixeira, Julia Wurie, Maryanne Sharma, Matthew Fox, et al. 2017. "A dedicated neonatal brain imaging system." *Magnetic Resonance in Medicine* 78.

Iglesias, Juan, Benjamin Billot, Yaël Balbastre, Azadeh Tabari, John Conklin, Daniel Alexander, Polina Golland, Brian Edlow, and Bruce Fischl. 2020. "Joint super-resolution and synthesis of 1 mm isotropic MP-RAGE volumes from clinical MRI exams with scans of different orientation, resolution and contrast." *ArXiv.*

Ioffe, Sergey, and Christian Szegedy. 2015. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift}." *Proceedings of the 32nd International Conference on Machine Learning* 448--456.

Isensee, Fabian, Paul Jaeger, Simon Kohl, Jens Petersen, and Klaus Maier-Hein. 2021. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." *Nature Methods* 1-9.

Jack, Clifford, Matt Bernstein, Nick Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, et al. 2008. "The Alzheimer's Disease neuroimaging initiative (ADNI): MRI methods." *Journal of magnetic resonance imaging : JMRI* 27: 685/691.

Jovicich, Jorge, Silvester Czanner, Douglas Greve, Elizabeth Haley, Andre Kouwe, Randy Gollub, David Kennedy, et al. 2006. "Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data." *NeuroImage* 436-43.

Jungo, Alain, Richard McKinley, Raphael Meier, Urspeter Knecht, Luis Vera, Julian Perez-Beteta, David Molina-Garcia, Victor M. Perez-Garcia, Roland Wiest, and Mauricio Reyes. 2018. "Towards Uncertainty-Assisted Brain Tumor Segmentation and Survival Prediction." *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2017* 474--485.

Kaissis, Georgios, Alexander Ziller, Jonathan Passerat-Palmbach, Théo Ryffel, Dmitrii Usynin, Andrew Trask, Ionésio Lima, et al. 2021. "End-to-end privacy preserving deep learning on multi-institutional medical imaging." *Nature Machine Intelligence.*

Kamnitsas, Konstantinos, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, et al. 2017. "Unsupervised Domain Adaptation in Brain Lesion Segmentation with Adversarial Networks." *Information Processing in Medical Imaging* 597-609.

Kamnitsas, Konstantinos, Christian Ledig, Virginia F.J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Daniel Rueckert, and Ben Glocker. 2017. "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation." *Medical Image Analysis* 36: 61-78.

Karlawish, Jason, Mark S. Cary, Jonathan Rubright, and Tom TenHave. 2008. "How redesigning AD clinical trials might increase study partners willingness to participate." *Neurology* 1883-1888.

Keane, Pearse, and Eric Topol. 2018. "With an eye to AI and autonomous diagnosis." *npj Digital Medicine.*

Kendall, Alex and Gal, Yarin. 2017. "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" *Proceedings of the 31st International Conference on Neural Information Processing Systems}* 5580–5590.

Khagi, Bijen, Chung Ghiu Lee, and Goo-Rak Kwon. 2018. "Alzheimer's disease Classification from Brain MRI based on transfer learning from CNN." *2018 11th Biomedical Engineering International Conference (BMEiCON)* 1-4.

Khosla, Meenakshi, Keith Jamison, Amy Kuceyeski, and Mert R. Sabuncu. 2019. "Ensemble learning with 3D convolutional neural networks for functional connectome-based prediction." *NeuroImage* 651-662.

Kornblith, Simon, Jonathon Shlens, and Quoc V. Le. 2019. "Do Better ImageNet Models Transfer Better?" *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks Systems." *Advances in Neural Information* 25.

Krizhevsky, Alex, Vinod Nair, and Geoffrey Hinton. 2009. "CIFAR-10 (Canadian Institute for Advanced Research)." *Technical Repor.*

Kushibar, Kaisar, Sergi Valverde, Sandra González-Villà, Jose Bernal, Mariano Cabezas, Arnau Oliver, and Xavier Llado. 2019. "Supervised Domain Adaptation for Automatic Sub-cortical Brain Structure Segmentation with Minimal User Interaction." *Scientific Reports* 9.

LeCun, Yann, Bernhard Boser, John Denker, Donnie Henderson, R. Howard, Wayne Hubbard, and Lawrence Jackel. 1990. "Handwritten Digit Recognition with a Back-Propagation Network." *Advances in Neural Information Processing Systems.*

LeCun, Yann, John S. Denker, and Sara A. Solla. 1990. "Optimal Brain Damage." *Advances in Neural Information Processing Systems 2* 598–605.

LeCun, Yann, Léon Bottou, Yoshi Bengio, and Patrick Haffner. 1998. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86: 2278-2324.

LeCun, Yann, Y. Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521: 436-44.

Lee, Bumshik, Nagaraj Yamanakkanavar, and Jae Young Choi. 2020. "Automatic segmentation of brain {MRI} using a novel patch-wise U-Net deep architecture." *PLOS ONE* 15: 1-20.

Lee, Hsin-Ying, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. 2020. "DRIT++: Diverse Image-to-Image Translation via Disentangled Representations." *International Journal of Computer Vision.*

Li, Hao, Asim Kadav, Igor Durdanovic, Hanan Samet, and H.P Graf. 2017. "Pruning Filters for Efficient ConvNets." *International Conference on Learning Representations.*

Li, Tian, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. "Federated Learning: Challenges, Methods, and Future Directions." *IEEE Signal Processing Magazine* 50-60.

Liu, Manhua, Fan Li, Hao Yan, Kundong Wang, Yixin Ma, Li Shen, and Mingqing Xu. 2020. "A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease." *NeuroImage* 116459.

Livne, Michelle, Jana Rieger, Orhun Utku Aydin, Abdel Aziz Taha, Ela Marie Akay, Tabea Kossen, Jan Sobesky, et al. 2019. "A U-Net Deep Learning Framework for High

Performance Vessel Segmentation in Patients With Cerebrovascular Disease." *Frontiers in Neuroscience* 13: 97.

Lu, Bin, Hui-Xian Li, Zhi-Kai Chang, Le Li, Ning-Xuan Chen, Zhi-Chen Zhu, Hui-Xia Zhou, et al. 2021. "A Practical Alzheimer Disease Classifier via Brain Imaging-Based Deep Learning on 85,721 Samples." *ArXiv.*

Luo, Xiangde, Guotai Wang, Tao Song, Jingyang Zhang, Michael Aertsen, Jan Deprest, Sebastien Ourselin, Tom Vercauteren, and Shaoting Zhang. 2021. "MIDeepSeg: Minimally Interactive Segmentation of Unseen Objects from Medical Images Using Deep Learning." *Medical Image Analysis* 102102.

Lyu, Ilwoo, Shuxin Bao, Lingya Hao, Jeweli Yao, Jacob A. Miller, Willa Voorhies, Warren D. Taylor, Silvia A. Bunge, Kevin S. Weiner, and Bennett A. Landman. 2021. "Labeling lateral prefrontal sulci using spherical data augmentation and context-aware training." *NeuroImage* 117758.

Madan, Christopher. 2021. "Scan Once, Analyse Many: Using Large Open-Access Neuroimaging Datasets to Understand the Brain." *Neuroinformatics.*

Maier-Hein, Lena, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, et al. 2018. "Why rankings of biomedical image analysis competitions should be interpreted with care." *Nature Communications.*

Major, David, Dimitrios Lenis, Maria Wimmer, Gert Sluiter, Astrid Berg, and Katja Bühler. 2020. "Interpreting Medical Image Classifiers by Optimization Based Counterfactual Impact Analysis." *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* 1096-1100.

Manjon, Jose. 2017. "MRI Preprocessing." *Imaging Biomarkers: Development and Clinical Integration* 53--63.

Marcus, Daniel, Tracy Wang, Jamie Parker, John Csernansky, John Morris, and Randy Buckner. 2007. "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults." *Journal of cognitive neuroscience* 19: 1498-1507.

Marek, Scott, Brenden Tervo-Clemmens, Ashley Nielsen, Muriah Wheelock, Ryland Miller, Timothy Laumann, Eric Earl, et al. 2019. "Identifying Reproducible Individual Differences in Childhood Functional Brain Networks: An {ABCD} Study." *Developmental Cognitive Neuroscience* 40: 100706.

McEver, R., and B. Manjunath. 2020. "PCAMs: Weakly Supervised Semantic Segmentation Using Point Supervision." *ArXiv.*

Mehmood, Atif, Shuyuan Yang, Zhixi Feng, Min Wang, AL Smadi Ahmad, Rizwan Khan, Muazzam Maqsood, and Muhammad Yaqub. 2021. "A Transfer Learning Approach for Early Diagnosis of Alzheimer's Disease on MR Images." *Neuroscience* 43-52.

Menze, Bjoern, András Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahaniy, Justin Kirby, Yuliya Burren, et al. 2014. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)." *IEEE Transactions on Medical Imaging* 99.

Mirza, Mehdi, and Simon Osindero. 2014. "Conditional Generative Adversarial Nets." *ArXiv.*

Mitra, Somosmita, Subhashis Banerjee, and Yoichi Hayashi. 2017. "Volumetric brain tumour detection from MRI using visual saliency." *PLOS One* 1-14.

Molchanov, Pavlo, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. 2016. "Pruning Convolutional Neural Networks for Resource Efficient Transfer Learning." *International Conference on Learning Representations.*

Morid, Mohammad Amin, Alireza Borjali, and Guilherme Del Fiol. 2021. "A scoping review of transfer learning research on medical image analysis using ImageNet." *Computers in Biology and Medicine* 128: 104-115.

Moyer, Daniel, Greg Ver Steeg, Chantal Tax, and Paul Thompson. 2020. "Scanner invariant representations for diffusion MRI harmonization." *Magnetic Resonance in Medicine.*

Murugesan, Balamurali, Sricharan Vijayarangan, Kaushik Sarveswaran, Keerthi Ram, and Mohanasankar Sivaprakasam. 2020. "KD-MRI: A knowledge distillation framework for image reconstruction and image restoration in MRI workflow." *Proceedings of the Third Conference on Medical Imaging with Deep Learning* 515--526.

Nath, Vishwesh, Dong Yang, Bennett A. Landman, Daguang Xu, and Holger R. Roth. 2021. "Diminishing Uncertainty within the Training Pool: Active Learning for Medical Image Segmentation." *ArXiv.*

Nguyen, Kevin P., Cherise Chin Fatt, Alex Treacher, Cooper Mellema, Madhukar H. Trivedi, and Albert Montillo. 2020. "Anatomically informed data augmentation for functional {MRI} with applications to deep learning." *Medical Imaging 2020: Image Processing* 172-177.

Ning, Lipen, Elisenda Bonet-Carne, Francesco Grussu, Farshid Sepehrband, Kaden Enrico, Jelle Veraart, Stefano B. Blumberg, et al. 2020. "Cross-scanner and cross-protocol multi-shell diffusion {MRI} data harmonization: Algorithms and results." *NeuroImage* 117128.

Nobis, Lisa, Stephen, Smith, Fidel Alfaro-Almagro, Mark Jenkinson, Clare Mackay, and Masud Husian. 2019. "Hippocampal volume across age: Nomograms derived from over 19,700 people in UK Biobank." *NeuroImage: Clinical* 101904.

Nooner, Kate, Stanley Colcombe, Russell Tobe, Maarten Mennes, Melissa Benedict, Alexis Moreno, Laura Panek, et al. 2012. "The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry." *Frontiers in neuroscience* 6.

Olah, Chris, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. "The Building Blocks of Interpretability." *Distill.*

Papernot, Nicolas, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2017. "Practical Black-Box Attacks against Machine Learning." *Association for Computing Machinery 2017* 506-519.

Park, JongChan, Sanghyun Woo, Joon-Young Lee, and Inso Kweon. 2018. "BAM: Bottleneck Attention Module." *Proceedings of the British Machine Vision Conference (BMVC) 2018.*

Paulus, Jessica, and David Kent. 2020. "Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities." *npj Digital Medicine* 99.

Pawlowski, Nick, Daniel Coelho de Castro, and Ben Glocker. 2020. "Deep Structural Causal Models for Tractable Counterfactual Inference." *Advances in Neural Information Processing Systems* 857--869.

Pawlowski, Nick, Daniel Coelho de Castro, and Ben Glocker. 2020. "Deep Structural Causal Models for Tractable Counterfactual Inference}." *Advances in Neural Information Processing Systems* 857--869.

Peng, Han, Weikang Gong, Christian F Beckmann, Andrea Vedaldi, and Stephen M Smith. 2021. "Accurate brain age prediction with lightweight deep neural networks." *Medical Image Analysis* 101871.

Perez-Garcia, Fernando, Rachel Sparks, and Sebastien Ourselin. 2020. "TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning." *ArXiv.*

Perone, Christian S., Pedro Ballester, Rodrigo C. Barros, and Julien Cohen-Adad. 2019. "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling." *NeuroImage* 1-11.

Radosavovic, Ilija, Piotr Dollar, Ross Girshick, Georgia Gkioxari, and Kaiming He. 2018. "Data Distillation: Towards Omni-Supervised Learning." *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 4119-4128.

Raghu, Maithra, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. 2019. "Transfusion: Understanding Transfer Learning with Applications to Medical Imaging." *33rd Conference on Neural Information Processing Systems (NeurIPS 2019).*

Rajchl, Martin, Matthew C. H. Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Mellisa Damodaram, et al. 2017. "DeepCut: Object Segmentation From Bounding Box Annotations Using Convolutional Neural Networks." *IEEE Transactions on Medical Imaging* 674-683.

Ramprasaath, R. Selvaraju, Michael Cogswell, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." *2017 IEEE International Conference on Computer Vision (ICCV)* 618-626.

Rieke, Nicola, Jonny Hancox, Fausto Milletari, Holger Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu Galtier, et al. 2020. "The Future of Digital Health with Federated Learning." *npj Digital Medicine volume.*

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* 234-241.

Roy, Abhijit Guha and Conjeti, Sailesh, Nassir Navab, and Christian Wachinger. 2018. "Inherent Brain Segmentation Quality Control from Fully ConvNet Monte Carlo Sampling." *Medical Image Computing and Computer Assisted Intervention -- MICCAI 2018* 664--672.

Rudin, Cynthia. 2019. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence.*

Schlemper, Jo, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. 2019. "Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images." *Medical Image Analysis.*

Shah, Nigam H., Arnold Milstein, and Steven C. Bagley. 2019. "Making Machine Learning Models Clinically Useful." *JAMA* 1351-1352.

Sheller, Micah J., G. Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. 2019. "Multi-institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation." *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* 92-104.

Shokri, Reza, and Vitaly Shmatikov. 2015. "Privacy-preserving deep learning." *2015 53rd Annual Allerton Conference on Communication, Control, and Computing* 909-910.

Shu, Michelle, Chenxi Liu, Weichao Qiu, and Alan Yuille. 2020. "Identifying Model Weakness with Adversarial Examiner." *Proceedings of the AAAI Conference on Artificial Intelligence* 11998-12006.

Shwartzman, Or, Harel Gazit, Ilan Shelef, and Tammy Riklin-Raviv. 2020. "The Worrisome Impact of an Inter-rater Bias on Neural Network Training." *ArXiv.*

Simard, Patrice, Yann Lecun, and John Denker. 1998. "Transformation invariance in pattern recognition tangent distance and tangent propagation." *Neural networks: tricks of the trade.*

Simonyan, Karen, and Andrew Zisserman. 2015. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *International Conference on Learning Representations.*

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2014. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." *Workshop at International Conference on Learning Representations.*

Singla, Sumedha, Brian Pollack, Stephen Wallace, and Kayhan Batmanghelich. 2021. "Explaining the Black-box Smoothly- A Counterfactual Approach." *ArXiv.*

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research* 15 (56): 1929-1958.

St-Jean, Samuel, Max Viergever, and Alexander Leemans. 2019. "Harmonization of diffusion MRI datasets with adaptive dictionary learning." *bioRXiv.*

Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, et al. 2015. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age." *PLOS Medicine* 12.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. "Axiomatic Attribution for Deep Networks." *Proceedings of the 34th International Conference on Machine Learning .*

Sundaresan, Vaanathi, Giovanna Zamboni, Nicola K. Dinsdale, Peter M. Rothwell, Ludovica Griffanti, and Mark Jenkinson. 2021. "Comparison of domain adaptation techniques for white matter hyperintensity segmentation in brain MR images." *ArXiv.*

Svensson, Carl, Ron Hübler, and Marc Figge. 2015. "Automated Classification of Circulating Tumor Cells and the Impact of Interobsever Variability on Classifier Training and Performance." *Journal of Immunology Research* 1-9.

Sweeney, Latanya. 2002. "K-Anonymity: A Model for Protecting Privacy." *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 557–570.

Tajbakhsh, Nima, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. 2020. "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation." *Medical Image Analysis* 101693.

Takacs, Petra, and Andrea Manno-Kovacs. 2018. "MRI Brain Tumor Segmentation Combining Saliency and Convolutional Network Features." *2018 International Conference on Content-Based Multimedia Indexing (CBMI)* 1-6.

Takao, Hidemasa, Naoto Hayashi, and Kuni Ohtomo. 2011. "Effect of Scanner in Longitudinal Studies of Brain Volume Changes." *Journal of magnetic resonance imaging : JMRI* 438-444.

Takao, Hidemasa, Naoto Hayashi, and Kuni Ohtomo. 2013. "Effects of Study Design in Multi-scanner Voxel-based Morphometry Studies." *Neuroimage.*

Talo, Muhammed, Ulas Baloglu, Ozal yıldırım, and U Rajendra Acharya. 2018. "Application of Deep Transfer Learning for Automated Brain Abnormality Classification Using MR mages." *Cognitive Systems Research.*

Tanner, Adam. 2017. *Our Bodies, Our Data: How Companies Make Billions Selling Our Medical Records.* Beacon Press.

Taylor, Jason R, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A. Shafto, Marie Dixon, Lorraine K. Tyler, Cam-CAN, and Richard N. Henson. 2017. "The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample." *Neuroimage* 144: 262-269.

Thagaard, Jeppe, Soren Hauberg, Bert van der Vegt, Thomas Ebstrup, Johan D. Hansen, and Anders B. Dahl. 2020. "Can You Trust Predictive Uncertainty Under Real Dataset Shifts in Digital Pathology?" *Medical Image Computing and Computer Assisted Intervention -- MICCAI 2020* 824--833.

Thakur, Siddhesh, Jimit Doshi, Sarthak Pati, Saima Rathore, Chiharu Sako, Michel Bilello, Sung Ha, et al. 2020. "Brain Extraction on MRI Scans in Presence of Diffuse Glioma: Multi-institutional Performance Evaluation of Deep Learning Methods and Robust Modality-Agnostic Training." *NeuroImage* 117081.

Tousignant, Adrian, Paul Lemaitre, Doina Precup, Douglas L. Arnold, and Tal Arbel. 2019. "Prediction of Disease Progression in Multiple Sclerosis Patients using Deep Learning Analysis of MRI Data." *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning* 483--492.

Tzeng, Eric, Judy Hoffman, Trevor Darrell, and Kate Saenko. 2015. "Simultaneous Deep Transfer Across Domains and Tasks." *2015 IEEE International Conference on Computer Vision (ICCV)* 4068-4076.

Valverde, Sergi, Mostafa Salem, Mariano Cabezas, Deborah Pareto, Joan C. Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Joaquim Salvi, Arnau Oliver, and Xavier Lladó. 2019. "One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks." *NeuroImage: Clinical* 101638.

Van Essen, David, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, and Kamil Ugurbil. 2013. "The WU_Minn Human Connectome Project: An overview." *Neuroimage* 80: 62-79.

Vaze, Sagar, Weidi Xie, and Ana I. L. Namburete. 2020. "Low-Memory CNNs Enabling Real-Time Ultrasound Segmentation Towards Mobile Deployment." *IEEE Journal of Biomedical and Health Informatics* 1059-1069.

Venturini, Lorenzo, Aris T. Papageorghiou, J. Alison Noble, and Ana I. L. Namburete. 2020. "Uncertainty Estimates as Data Selection Criteria to Boost Omni-Supervised Learning." *Medical Image Computing and Computer Assisted Intervention -- MICCAI 2020* 689--698.

Verma, Sahil, John Dickerson, and Keegan Hines. 2020. "Counterfactual Explanations for Machine Learning: A Review." *ArXiv.*

Vidanage, Anushka, Peter Christen, Thilina Ranbaduge, and Rainer Schnell. 2020. "A Graph Matching Attack on Privacy-Preserving Record Linkage." *Proceedings of the 29th ACM International Conference on Information and Knowledge Management* 1485-1494.

Visser, Martin, Domenique Müller, R.J.M. Duijn, M. Smits, Niels Verburg, Eef Hendriks, R.J.A. Nabuurs, et al. 2019. "Inter-rater agreement in glioma segmentations on longitudinal MRI." *NeuroImage: Clinical* 101727.

Wachinger, Christian, Martin Reuter Reuter, and Tassilo Klein. 2018. "DeepNAT: Deep convolutional neural network for segmenting neuroanatomy." *NeuroImage* 170: 434-445.

Wang, Guotai, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. 2019. "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks." *Neurocomputing* 34-45.

Wang, Lu, Dong Guo, Guotai Wang, and Shaoting Zhang. 2020. "Annotation-Efficient Learning for Medical Image Segmentation based on Noisy Pseudo Labels and Adversarial Learning." *IEEE Transactions on Medical Imaging.*

Wang, X., R. Girshick, A. Gupta, and K He. 2018. "Non-local Neural Networks." *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 7794-7803.

Wang, Zhibo, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. 2019. "Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning." *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications* 2512-2520.

Williams, David R., and Ronald Wyatt. 2015. "Racial Bias in Health Care and Health: Challenges and Opportunities." *JAMA* 555-556.

Winkler, Julia K., Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, et al. 2019. "Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition." *JAMA Dermatology* 1135-1141.

Wong, Winston, Thomas Laveist, and Joshua M. Sharfstein. 2015. "Achieving Health Equity by Design." *JAMA.*

Wu, Wenshan, Yuhao Lu, Ravikiran Mane, and Cunta Guan. 2020. "Deep Learning for Neuroimaging Segmentation with a Novel Data Augmentation Strategy." *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)* 1516-1519.

Xia, Tian, Agisilaos Chartsias, Chengjia Wang, and Sotirios A. Tsaftaris. 2021. "Learning to synthesise the ageing brain without longitudinal data." *Medical Image Analysis.*

Yang, Lin, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. 2017. "Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation." *Medical Image Computing and Computer Assisted Intervention -- MICCAI 2017* 399--407.

Yang, Qiang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. "Federated Machine Learning: Concept and Applications." *ACM Trans. Intell. Syst. Technol.*

Yongchan, Kwon, Joong Ho Won, Beom Joon Kim, and Myunghee Cho Paik. 2020. "Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation." *Computational Statistics and Data Analysis.*

Yosinski, Jason, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. "How transferable are features in deep neural networks?" *Advances in Neural Information Processing Systems* 27.

Young, Alexandra, Razvan Marinescu, Neil Oxtoby, Martina Bocchetta, Keir Yong, Nicholas Firth, David Cash, et al. 2018. "Uncovering the heterogeneity and temporal

complexity of neurodegenerative diseases with Subtype and Stage Inference." *Nature Communications* 9.

Yu, Meichen, Kristin Linn, Philip Cook, Mary Phillips, Melvin McInnis, Maurizio Fava, Madhukar Trivedi, Myrna Weissman, Russell Shinohara, and Yvette Sheline. 2018. "Statistical harmonization corrects site effects in functional connectivity measurements from multisite fMRI data." *Human Brain Mapping* 39.

Zeiler, Matthew, and Rob Fergus. 2014. "Visualizing and Understanding Convolutional Neural Networks." *2014 IEEE European Conference on Computer Vision (ECCV).*

Zendle, Oliver, Markus Murschitz, Martin Humenberger, and Wolfgang Herzner. 2015. "CV-HAZOP: Introducing Test Data Validation for Computer Vision." *ICCV.*

Zhang, Hongyi, Moustapha Cisse, Yann Dauphin, and David Lopez-Paz. 2017. "mixup: Beyond Empirical Risk Minimization." *ArXiv.*

Zhang, Yuhao, Hang Jiang, Yasuhide Miura, Christopher Manning, and Curtis Langlotz. 2020. "Contrastive Learning of Medical Visual Representations from Paired Images and Text." *ArXiv.*

Zhou, Tongxue, Stéphane Canu, Pierre Vera, and Su Ruan. 2020. "Brain Tumor Segmentation with Missing Modalities via Latent Multi-source Correlation Representation." *Medical Image Computing and Computer Assisted Intervention -- MICCAI 2020* 533--541.

Zhou, Zongwei, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B. Gotway, and Jianming Liang. 2019. "Models Genesis: Generic Autodidactic Models for {3D} Medical Image Analysis." *Medical Image Computing and Computer Assisted Intervention -- MICCAI 2019* 384-393.

Zhu, Hancan, Zhenyu Tang, Hewei Cheng, Yihong Wu, and Yong Fan. 2019. "Multi-atlas label fusion with random local binary pattern features: Application to hippocampus segmentation." *Scientific Reports* 16839.

Zhuang, Peiye, Alexander G. Schwing, and Oluwasanmi Koyejo. 2019. "FMRI Data Augmentation Via Synthesis." *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* 1783-1787.

Zhuang, Xinrui, Yuexiang Li, Yifan Hu, Kai Ma, Yujiu Yang, and Yefeng Zheng. 2019. "Self-supervised Feature Learning for 3D Medical Images by Playing a Rubik's Cube." *Medical Image Computing and Computer Assisted Intervention -- MICCAI 2019* 420-428.

Zhuang, Zhemin ; Li, Nan ; Joseph, Raj; Alex, Noel ; Mahesh, Vijayalakshmi ; Qiu, Shunmin. 2019. "An RDAU-NET model for lesion segmentation in breast ultrasound images." *PLOS One.*

Ziller, Alexander, Dmitrii Usynin, Rickmer Braren, Marcus Makowski, Daniel Reuckert, and George Kaissis. 2021. "Medical imaging deep learning with differential privacy." *Nature Scientific Reports.*

Zitnick, C.W., N. Jojic, and Bing Kang Sing. 2005. "Consistent segmentation for optical flow estimation." *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1* 2: 1308-1315.

Zuo, Xi-Nian, Jeffrey Anderson, Pierre Bellec, Rasmus Birn, Bharat Biswal, Janusch Blautzik, John Breitner, et al. 2015. "An open science resource for establishing reliability and reproducibility in functional connectomics." *Scientific Data* 1.