From YouTube to the Brain: Transfer Learning Can Improve Brain-Imaging Predictions with Deep Learning

Nahiyan Malik

MASTER OF SCIENCE

School of Computer Science

McGill University

Montreal, Quebec, Canada

October 2021

A thesis submitted to McGill University in partial fulfillment

of the requirements of the degree of Master of Science

Abstract

Deep learning has recently achieved best-in-class performance in several fields, including biomedical domains such as X-ray images. Yet, data scarcity poses a strict limit on training successful deep learning systems in many, if not most, biomedical applications, including those involving brain images. In this study, we translate state-of-the-art transfer learning techniques for single-subject prediction of simpler (sex and age) and more complex phenotypes (number of people in household, household income, fluid intelligence and smoking behavior). We fine-tuned 2D and 3D ResNet-18 convolutional neural networks for target phenotype predictions from brain images of ~40,000 UK Biobank participants, after pretraining on Youtube videos from the Kinetics dataset and natural images from the ImageNet dataset. Transfer learning was effective on several phenotypes, especially sex and age classification. Additionally, transfer learning in particular outperformed deep learning models trained from scratch especially on smaller sample sizes. The out-of-sample performance using transfer learning from previously learned knowledge based on real world images and videos could unlock the potential in many areas of imaging neuroscience where deep learning solutions are currently infeasible.

Abrégé

L'apprentissage en profondeur a récemment atteint les meilleures performances de sa catégorie dans plusieurs domaines, notamment des domaines biomédicaux tels que les images radiographiques. Pourtant, la rareté des données pose une limite stricte à la formation de systèmes d'apprentissage en profondeur réussis dans de nombreuses, sinon la plupart, des applications biomédicales, y compris celles impliquant des images cérébrales. Dans cette étude, nous traduisons des techniques d'apprentissage par transfert de pointe pour la prédiction sur un seul sujet de phénotypes plus simples (sexe et âge) et plus complexes (nombre de personnes dans le ménage, revenu du ménage, intelligence fluide et comportement tabagique). Nous avons affiné les réseaux de neurones convolutifs ResNet-18 2D et 3D pour les prédictions de phénotype cible à partir d'images cérébrales d'environ 40 000 participants à la biobanque britannique, après une pré-formation sur des vidéos Youtube de l'ensemble de données Kinetics et des images naturelles de l'ensemble de données ImageNet. L'apprentissage par transfert a été efficace sur plusieurs phénotypes, en particulier la classification du sexe et de l'âge. De plus, l'apprentissage par transfert en particulier a surpassé les modèles d'apprentissage en profondeur formés à partir de zéro, en particulier sur des échantillons de plus petite taille. Les performances hors échantillon utilisant l'apprentissage par transfert à partir de connaissances acquises précédemment basées sur des images et des vidéos du monde réel pourraient libérer le potentiel dans de nombreux domaines de la neuroscience de l'imagerie où les solutions d'apprentissage en profondeur sont actuellement infaisables.

Acknowledgements

I would like to thank Prof. Danilo Bzdok for all of his guidance and help throughout my Master's. His intuition and leadership were invaluable not only for this work, but for my future endeavours as well. This research was possible due to the funding provided by Prof. Bzdok. The duration of my degree was challenging due to the COVID-19 pandemic, which hindered the collaboration that I sought out when I moved to Montreal for my studies. However, constantly being connected via virtual meetings made the best of the situation and I really appreciated the abundance of time that was allocated for me by Prof. Bzdok.

I would also like to thank Blake Richards, Pouya Bashivan, Guillaume Lajoie and Irina Rish for suggestions on an earlier version of this work. This work would also not be possible without the UK Biobank dataset. Such large-scale datasets enable novel studies of this kind.

I would be remiss to not mention some of the earliest lab members: Emile, Hasnain and Hannah. Unfortunately, we did not have the opportunities to meet and collaborate much in person, but we always kept in touch virtually which was a big help and morale booster.

Finally, I would not be who I am today without my family. In many ways, I'm not supposed to be here. Their constant struggle and support have paved for me roads I still cannot comprehend. To them I dedicate everything.

Contribution of Authors

All of the work in this research including the writing of the thesis was conducted by the student.

Feedback and guidance were provided by the supervisor.

Table of Contents

1. Introduction **5**

2. Literature Review **8**

2.1. Imaging neuroscience and machine learning 8

2.2. Deep learning 9

2.3. Transfer learning 11

2.4. Phenotypes split by sex 13

3. Methods **14**

3.1. Rationale and workflow summary 14

3.2. Population data resource 15

3.3. Target phenotypes 16

3.4. Model selection: cross validation schemes for linear baseline and DNNs 18

3.5. Linear baseline 20

3.6. Transfer learning techniques 20

3.7. Deep learning models 25

4. Results **30**

4.1. Sex classification 30

4.2. Classifying more complex phenotypes 34

5. Discussion **50**

6. Conclusion **58**

References **59**

1. Introduction

The advent of deep learning has brought forth transformative results in domains such as image and video classification, natural language processing and audio recognition (LeCun et al., 2015). Similarly, medical imaging has also seen state-of-the-art success through the application of deep learning (Biswas et al., 2019). However, deep neural networks (DNN) need to ingest a large amount of data to achieve effective predictions which are difficult to attain in many fields including brain imaging. As a means to a possible turning point, the UK Biobank (UKBB) is the largest currently existing biomedical dataset in the world. This resource also provides images from various brain imaging modalities of ~40,000 participants, with the anticipated goal of acquiring data from ~100,000 participants by 2022 (Miller et al., 2016). Previous research has investigated the application of various machine learning models on brain imaging data from the UKBB where DNNs often struggled to outperform simple linear models (Schulz et al., 2020). These benchmarking analyses indicated that much larger datasets could be required for DNNs to be able to fully exploit the non-linearities in the brain images to outperform even standard linear models. As such, the challenge is to effectively train DNNs on brain imaging data despite its data scarcity.

Transfer learning is the notion of using previously learned knowledge to aid in the learning of a new and different task from another dataset. More specifically, a DNN that has been trained on a larger and more general purpose dataset for a base task can then be fine-tuned for a wholly different target task (Yosinski et al., 2014). A DNN is first trained on the base task using an original dataset such as Kinetics to identify hierarchical non-linear representations in the features. In contrast, the target task has a much smaller dataset and is more specific such as a brain imaging classification task that is used to adapt the learned weights from the base task. Although DNNs trained on videos from the Kinetics dataset or natural images from the

ImageNet dataset learn to classify different tasks, the DNNs learn a non-linear cascade of processing operations to aid in its overall generalization. For example, the lowest and most general layers of a DNN typically detect features such as edges and curves that have been reported to be universally useful in most image recognition tasks. In comparison to real life, humans develop their vision through sights and sounds of real world experiences that in turn help with more specific tasks over time. In this sense, transfer learning can be seen as the acquisition of knowledge from a vast array of images and videos from a base task which in turn can be fine-tuned to be used on a target task that may not be strictly related to the original base task. There have been successful uses of transfer learning in other biomedical based machine learning studies including diagnosis of appendicitis, analysis of abdominal images and detection of Alzheimer's Disease (Cheng & Malhi, 2017; Khan et al., 2019; Rajpurkar et al., 2020). By means of transfer learning, it may be possible to take advantage of the massive amounts of general knowledge based on general-purpose real-world images and actions to then fine-tune it to our brain imaging classification tasks. In fact, bringing transfer learning to the brain imaging community may be one of the best avenues to enable training of successful DNNs in areas that suffer from data scarcity.

In the biology of the brain, sex differences have been found to be one of the most important sources of variability (Ritchie et al., 2018). There have been observed sex differences in previous studies focused on topics ranging from daily social lifestyle to psychiatric disorders (Iraji et al., 2021; Kiesow et al., 2020). In order acknowledge and exploit the many distinctive features found in the brains of males and female, we incorporate sex difference into our machine learning pipelines. Additionally, phenotype targets as well as machine learning models of varying complexities may exhibit different results. Therefore, it is important to investigate various transfer learning techniques on 2D and 3D DNNs in a systematic manner to observe any effects. Our goal thus in this study is to combine the UKBB brain imaging dataset with novel

transfer learning techniques to take advantages of learned knowledge from large, general purpose datasets to effectively train DNNs based on data split by sex.

2. Literature Review

2.1. Imaging neuroscience and machine learning

Biomedical domains have seen shifts in datasets allowing for more complex data analysis using varying levels of machine learning in recent years (Foster et al., 2014). Use cases such as classification of arrhythmia from ECG signals is one of many examples of using machine learning on biomedical datasets. However, challenges have needed to be overcome to be able to successfully apply machine learning to such datasets: the limitations of sample sizes as well as selecting and validating models appropriate for the datasets. Many such datasets offer limited amounts of structured and labeled data which inherently places constraints on models that are fit for the tasks. Models such as logistic regression and support vector machines (SVM) have been the standard for analyses due to the lower computation requirements as well as smaller sample sizes (Bzdok, 2017; Bzdok & Ioannidis, 2019; Efron & Hastie, 2016). Additionally, data analysis tasks in the biomedical domain also often require inference and association to better understand the effects of input variables on the classification results. However, the focus put on inference compared to prediction is solely based on the overall goal of the analysis (Bzdok et al., 2020; Bzdok & Ioannidis, 2019). There is also a growing focus on personalized medicine along with increasing sample sizes. In such settings where the outcomes may be uncommon or unobserved based on specific individuals, raw prediction performance could be key (Bzdok et al., 2021).

Imaging neuroscience has recently experienced the arrival of large datasets such as the UKBB with multiple modalities and high resolution images (Sudlow et al., 2015). Due to the larger sample sizes as well as the vast array of target variables that are included in the UKBB, novel analyses that were not previously possible are now able to be investigated. Not only does the

UKBB aim to have ~100,000 participants' brain imaging data by the year 2022, the scope in addition to the size is impressive. Each participant also has lifestyle outcomes, biological measurements, brain and body imaging as well as biomarkers. There are also follow-up data for participants to enable longitudinal studies (Bycroft et al., 2018). Naturally, a dataset of such scale and scope warrants experiments that include more complex, non-linear machine learning models with potentially greater prediction capabilities.

A previous study carried out a set of thorough benchmarking analyses on the UKBB across various machine learning models on brain imaging data (Schulz et al., 2020). It was found that even while using MRI brain scans from the UKBB dataset, deep learning struggled to outperform simpler linear models. The authors posit the main reason for this is data scarcity. That is, even using the full UKBB dataset at the time, there may not be enough samples to effectively train DNNs and that larger amounts of training data would be needed for better applications of deep learning in imaging neuroscience. Similarly, He et al. had found that DNNs were not able to outperform kernel regression models on functional connectivity data to predict fluid intelligence (T. He et al., 2018). On the same note as Schulz et al., the authors state that more samples may be required to apply deep learning to brain imaging.

2.2. Deep learning

Deep learning has become the state-of-the-art for many machine learning tasks. DNNs encompass hierarchical non-linearities across many hidden layers, starting from the input variables until the outputs. Whereas traditional machine learning techniques require manual feature engineering of the input data, a DNN learns by itself the best feature representations through its multi-level abstraction (Goodfellow et al., 2016). For example, in terms of an image classification task, the early, lower layers of a DNN would typically learn basic shapes like

edges, curves and color blobs. The later layers would build upon the hierarchy of the lower layers to form together more specific shapes such as the tires of a car. The latest, highest layers would then combine together the abstracted information learned from the previous layers to classify images such as cars. DNNs in turn have produced state of the art results in tasks such as image and video recognition, speech and audio recognition and natural language processing (LeCun et al. 2015).

Convolutional neural networks (CNN) in particular have had tremendous success in image processing. In 2012, AlexNet won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), sparking the vast advancements in this kind of technology (Krizhevsky et al., 2012). Since then, there have been various types of CNN architectures providing iterative improvements such as VGGNet, GoogleLeNet and ResNet (K. He et al., 2016; Simonyan & Zisserman, 2014; Szegedy et al., 2014). The main goal for all these architectures has been to improve overall prediction performance with some opting for more layers while others trying to be more efficient. Generally, it has been found that deeper networks perform the best. ResNet, for example, implements residual blocks to avoid the vanishing gradient problem while keeping a greater number of layers.

The success of deep learning has also been ever increasing in medical imaging. In a meta-analysis by Aggarwal et al. deep learning resulted in impressive classification accuracies (Aggarwal et al., 2021). Opthalmology tasks for diagnosis of diabetic retinopathy, age-related macular degeneration and glaucoma from retinal fundus images and optical coherence tomography ranged between AUCs of 0.933 and 1. Respiratory imaging tasks for diagnosis of lung nodules in lung cancer using chest X-rays or CT scans ranged from AUCs of 0.864 and 0.937. Breast imaging diagnosis from mammogram, ultrasound, MRI and digital breast tomosynthesis had AUCs ranging from 0.868 and 0.909. In addition to just classification

accuracy, Liu et al. have compared the performance of deep learning classification performance to health-care professionals when classifying diseases from medical imaging (Liu et al., 2019). In their meta-analysis across studies filtering for highest accuracy, they found deep learning models to have a pooled sensitivity of 87.0% (95% CI 83.0-90.2) compared to 86.4% (95% CI 79.9-91.0) by health-care professionals. In terms of specificity, the deep learning models had a pooled specificity of 92.5% (95% CI 85.1-96.4) compared to 90.5% (95% CI 80.6-95.7) by health-care professionals. However, as Schulz et al. have found, deep learning struggles in the domain of imaging neuroscience due to data scarcity (Schulz et al., 2020). The lack of availability of enough data to train deep learning models effectively is not uncommon especially for more specialized tasks. Such models cannot solve any task just through the complexity of the models themselves; there needs to be a balance between the models and enough data required to utilize them (Emmert-Streib et al., 2020).

2.3. Transfer learning

In order to more efficiently use deep learning with smaller datasets, transfer learning shows potential. Although the ultimate goal for any given task could be to have sufficiently large sample sizes, it is oftentimes infeasible to do so due to the novelty and difficulty of acquiring more data. In such scenarios transfer learning can be applied, which uses knowledge learned from a larger source dataset to then modify and apply the information on a smaller target dataset by fine-tuning the weights of the DNN (Hutchinson et al., 2017; Yosinski et al., 2014). DNNs using transfer learning have been shown to be effective in many tasks (Weiss et al., 2016).

Image recognition has benefited greatly from transfer learning. Comparing the performance of models pretrained on ImageNet and models trained from scratch, Hossain et al. found transfer learning to provide much better classification accuracy on the CIFAR-10 dataset with 70%

compared to 38% when trained from scratch using a much smaller dataset (Hussain et al., 2019). Additionally, Zawadzka-Gosk et al. found transfer learning exceeded 99% classification accuracy on a car model recognition task (Zawadzka-Gosk et al., 2019). Video, which can be seen as a series of 2D image frames, has also seen performance improvements from utilizing transfer learning. Sarhan et al. found that first pretraining on the Kinetics dataset resulted in better than state-of-the-art accuracy on sign language recognition on the ChaLearn249 Isolated Gesture Recognition dataset. Audio processing has similarly benefitted from transfer learning. Various speech recognition tasks have shown to train faster when first pretrained on a source dataset indicating much more efficient training requiring less time. When fine-tuning is applied to the pretrained models, the classification accuracies were often found to be greater than with models trained from scratch (Kunze et al., 2017; Qin et al., 2018; Wang & Zheng, 2015). Even natural language processing has been positively impacted by transfer learning. Howard et al. show their ULMFiT model when using transfer learning vastly outperforms state of the art results in multiple classification tasks while requiring 100x less data (Howard & Ruder, 2018).

Although the tasks and types of datasets differ, the main idea of reusing knowledge that is acquired from a larger source dataset to be applied to a smaller target dataset is a common theme across transfer learning applications. Not only does transfer learning outperform many state of the art solutions, but the lower amount of training data is key for data-scare tasks such as many in the biomedical domain. There have been many notable and successful uses of transfer learning in the medical imaging field including tasks in kidney diagnosis, appendicitis detection, classification of Alzheimer's disease and abdominal ultrasound images (Cheng & Malhi, 2017; Khan et al., 2019; Rajpurkar et al., 2020; Ravishankar et al., 2017). However, as discussed by Schulz et al. brain imaging remains a challenge to effectively pair with deep learning due to its limited sample sizes (Schulz et al., 2020).

2.4. Phenotypes split by sex

Sex difference is one of the most important sources of variability in biology and the human brain (Ritchie et al., 2018). In order to more effectively classify different phenotypes using brain imaging data, we therefore also focused on sex differences in the brain. However, acknowledging and applying sex differences in data analysis tasks have not been very common. Sex plays a wide range of factors on human brain function (Cahill, 2014). Jazin et al. have shown that the sex dimorphisms in the brain stem from molecular neuroscience to ultimately modify signal pathways and thus need to be considered in studies where distinguishing between males and females is possible (Jazin & Cahill, 2010).

Previous research shows many instances of effects of sex difference. When accounting for sex difference, males and females have showcased varying anatomical measurements, differences in antisocial behaviour, different brain volumes from early age trauma, distinct theta activity during fluid intelligence tasks as well as social stimulation factors (Anderson et al., 2019; Badura-Brack et al., 2020; Kiesow et al., 2020; Ritchie et al., 2018; Taylor et al., 2020). There have also been noted differences between males and females in functional connectivity during response inhibition, processing of the visual network as well as connectivity in intrinsic brain dynamism (Cai et al., 2020; Chung et al., 2020; de Lacy et al., 2019). Due to sex difference being such a great source of variability in the human brain, all the following analyses include splitting of the phenotypes by male and female.

3. Methods

3.1. Rationale and workflow summary

To explore how to make the most out of the sample sizes available today, we tested for gains from transfer learning using 3D DNNs pretrained on Youtube videos from Kinetics and 2D DNNs pretrained on images from ImageNet. The obtained pretrained DNNs were then fine-tuned to our target classification tasks based on brain images. For training and fine-tuning the 2D DNNs, single image slices were extracted from the 3D structural brain scans along the sagittal plane in the anterior-posterior direction to use as input variables. In contrast, the 3D pretrained DNNs were fed “videos of brain images” as inputs. That is, we interpreted each 3D structural brain scan as an ordered series of 2D brain images along the sagittal plane in the anterior-posterior direction for a given participant. In so doing, the full collection of component brain images were used as input to our 3D DNNs. Using this feature engineering and model fine-tuning strategy, we have examined prediction performance of a range of simple to more complex target phenotypes. Doing so allowed for apples-to-apples comparisons across phenotypes and model complexities.

We further split up the phenotypes by sex, which was critical in order to observe differences in prediction accuracies for males and females. As a consequence, we systematically conducted our phenotype prediction experiments (Table 1) broken up by sex across linear and non-linear models of different representational capacities applied to the UKBB dataset. In particular, we trained a series of complementary deep learning architectures, starting with a linear baseline and progressing to 2D and 3D DNNs. We intended to observe the predictive capabilities of each model on simple and more complex phenotypes. Being the largest dataset of its kind, the UKBB

enables us to experiment with larger, more complex predictive models to investigate the differences in predicting phenotypes than was possible before.

3.2. Population data resource

The UKBB dataset is the largest uniformly acquired brain imaging dataset in the world. In 2014 the UKBB initiative started collection of its brain imaging supplement. This extension included acquisition of several brain MRI modalities from ~100,000 subjects by 2022 (Miller et al., 2016). This study used the available data as of February/March 2020. As part of the large dataset, various phenotype descriptors allow for a comprehensive investigation of classification tasks. We used T1-weighted MRI imaging measures of grey matter morphology (sMRI, structural magnetic resonance imaging) from 38,701 participants. We also examined the scenario where only brain scans from a smaller subset from 5,000 subjects are available, which recapitulates the sample size of the first UKBB imaging release (Miller et al., 2016). The smaller and larger sample sizes make it possible to observe the effects of scaling behavior on our classification tasks. All analyses in this study were conducted using the UKBiobank resource under Data Access Application 25163. All participants were informed and consented to participate. Further details about the consent procedures can be found online (https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=200).

3.3. Target phenotypes

The present investigation considered six target phenotypes: sex, age, number in household, household income, fluid intelligence and past smoking, following previous research (Schulz et al., 2020). Using phenotypes from past research enabled a level of comparison to past work. These phenotypes also range in expected challenge in extracting meaningful brain patterns for

the goal of out-of-sample prediction. The sex and age phenotypes explain a lot of the signal from the MRI-derived data. In contrast, the number in household, fluid intelligence, household income and past smoking phenotypes present much more challenging classification tasks.

Table 1: Target phenotypes examined in our study.

Phenotype	**UKBB Data-Field**	**Description**
Sex	31	Participant's sex at birth.
Number in household	709	Participant's answer to the question "Including yourself, how many people are living together in your household?" (https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=709)
Fluid intelligence	20016	Sum of the number of correct answers given to the 13 fluid intelligence questions, such as "add the following numbers together", "stop means the same as", and "relaxed means the opposite of" (https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=20016)
Household income	738	Participant's answer to the question, "What is

		the average total income before tax received by your household?" The answers are grouped into 7 categories: less than 18,000, 18,000 to 30,999, 31,000 to 51,999, 52,000 to 100,000, greater than 100,000, do not know and prefer not to answer. (https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=738)
Past smoking	1249	Participant's answer to the question "In the past, how often have you smoked tobacco?" The answers are grouped into 5 categories: smoked on most or all days, smoked occasionally, just tried once or twice, and I have never smoked. (https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=1249)
Age	21003	Participant's age at time of recruitment. (https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=21003)

For the large-sample scenario, sex classification included the whole ~40,000 subjects and the classification tasks for the rest of the target phenotype predictions included ~20,000 subjects. As an added layer of analysis in our study, we split all phenotype classification tasks by sex, which led to all phenotypes (except sex itself) to have approximately half of the total subject count. For example, the number in household phenotype was split into two classification tasks: number in household (male) and number in household (female).

3.4. Model selection: cross validation schemes for linear baseline and DNNs

All of the analyses in this study across all phenotypes and sample sizes consisted of a linear baseline using Linear Support Vector Classification (SVC) and non-linear deep neural networks that can identify and exploit hierarchical non-linear representations in the input variables. Each of the two types of model representation capacity conformed to a different cross validation scheme. Along with the Linear SVC, we used a nested k-fold cross validation for 10 outer loops and 5 inner loops using stratified shuffle splits. The stratified shuffle split first permuted the order of the participant data points and split it into training and test sets while keeping the class balance of the original dataset. The inner loop of the nested cross validation was used for hyperparameter tuning and model selection. Instead, the outer loop is used to obtain a principled estimate of the predictive accuracy that we would expect in fresh data points or newly recruited participants. This procedure allowed us to get an unbiased measure of the predictive ability of the model across multiple out-of-sample predictions. The described cross validation scheme was used for all linear baseline predictions across sample sizes and phenotypes.

In comparison to the linear baseline, training the DNNs required a considerably larger computational budget and took longer to complete. As a commonly practiced scheme in the deep learning community (Goodfellow et al., 2016), we used a random, but fixed partition of our

participant sample into an 80% training set, 10% validation set and 10% test set. In order to explicitly quantify the uncertainty attributable to sampling variation, we obtained model estimates across different model parameter initializations. Three different initializations were used per sample size and per target phenotype. The initialization of the DNN parameters was influenced by two sources of randomness: the initial random weights of the model and the order of the input data passed into the mini-batches in stochastic gradient descent. The seeds for the random generators across runs were kept consistent for future reproducibility. All other aspects of the data analysis settings were kept constant across the 3 different parameter initializations to ensure reproducibility of our results. For the pretrained models, fine-tuning was done on the training set. All hyperparameter tuning was done exclusively on the validation set. Instead, the participant data from the test set was solely used for out-of-sample prediction from the hyperparameter-tuned model. The learning rates to train from scratch as well as the discriminative learning rates for transfer learning are presented below (cf. Transfer learning). For all of the considered DNN models, we used the ADAM optimization algorithm as solver (Kingma & Ba, 2014), a batch size of 8 data points or participants and a maximum of 250 epochs with an early stopping criterion that took effect at 10 epochs with no decrease in validation loss.

3.5. Linear baseline

Support vector machines have been a widely used classification algorithm for biomedicine (Bishop, 2006; Murphy, 2012). In order to compare against other, more complex models as well as to understand the efficacy of a linear model on the dataset, a Linear SVC for binary classification was used as the linear baseline. The features of the Linear SVC were principal components of the MRI scans.

Using the Nilearn python library (https://nilearn.github.io), the brain scans were preprocessed by slicing and dicing before model fitting. The MRI scans were resampled to a dimension of 91x108x91 based on the MNI mask resolution. Then, a mask was fitted on the MRI scans to then transform each MRI scan using the predefined grey matter mask, which yielded a final resolution of 296,811 grey matter voxels. The voxel features were z-scored across all participants within the cross validation pipeline in the respective iterations followed by dimension reduction using PCA within the grid search used for hyperparameter tuning (cf. below). Nested cross validation was done with 10 outer loops and 5 inner loops using stratified shuffle splits. As part of the nested cross validation, grid search was used for hyperparameter tuning of the number of PCA components and the Linear SVC C regularization parameter. The grid search values for the PCA components were 100 and 500 and the Linear SVC C values that control the regularization strength as hyperparameter values were 0.1, 0.25, 1, 25 and 100. This combination of hyperparameters led to an exploration of 10 different modeling regimes that were probed and selected on the validation set.

3.6. Transfer learning techniques

DNNs have been shown to struggle to outperform linear models in certain brain imaging tasks, which could partially be due to the smaller dataset sizes as demonstrated in our previous work (Schulz et al., 2020). Transfer learning is a methodical toolkit that can potentially overcome some of these limitations of earlier use in deep learning in brain imaging data. In transfer learning, we can take the structured knowledge that a DNN has gleaned from one base task and apply it to a different target task. To make progress towards this goal, we have built on Kinetics and ImageNet as our large, general-purpose datasets (Deng et al., 2009; Kay et al., 2017).

Kinetics is one of the largest, openly available datasets that provides high-quality videos taken from Youtube. It consists of a range of everyday human actions. More specifically, the Kinetics-400 dataset is composed of 276,708 separate videos that were labeled to belong to one of 400 different human actions. Such action categories include breakdancing, hitting a baseball, windsurfing, eating spaghetti, riding a bike and so forth. For the purpose of the present study, we considered the video clips from the Kinetics dataset to be analogous to the 3D MRI brain scans as both are made up of a series of 2D frames. In addition to the Kinetics resource, we also gauged the performance of using transfer learning on 2D slices taken from the 3D MRI scans on the sagittal plane using the ImageNet dataset. ImageNet is one of the largest repositories of 2D images with over 14,000,000 images spanning more than 20,000 categories such as cats, dogs, helicopters, cars, apples, bananas, etc. It is one of the most widely used datasets in machine learning, highlighted by the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which has been pivotal in the progress of computer vision, especially convolutional neural networks (Russakovsky et al., 2015).

It is crucial when applying transfer learning to avoid catastrophic forgetting: the loss of knowledge acquired from the base task while adapting the pretrained weights for the target task (Kirkpatrick et al., 2017). The pretrained weights after training on the base task (i.e., distinguishing 400 actions in Youtube videos or 20,000 categories in natural images) encode the knowledge that is gained from the base dataset. The lowest layers (closest to the input) of such a model typically extract the most general features such as edges, curves, features similar to Gabor filters and color blobs (Yosinski et al., 2014). The layers get more and more specific to the actual prediction task at hand, with the highest layers (closest to the output) having the most specific features. The base model with the pretrained weights is then further trained using the target dataset which in our case is the UKBB brain scans. The training set of the target dataset further adjusts the pretrained weights to adapt to the target task; this process is known as

fine-tuning. Thus, the goal was to keep most of the learned general knowledge from the large, general-purpose dataset (Kinetics or ImageNet) that are in the lower layers and in comparison, update the specific knowledge in the higher layers more by modifying particular sets or layers of weights in order to avoid catastrophic forgetting. We have used discriminative learning rates (Figure 1, Table 2) to achieve this strategy (Howard & Ruder, 2018). By using discriminative learning rates, the learning rates are largest in the highest, most specific layers and are reduced for lower, more general layers, which in turn modifies the higher, specific layers more compared to the lower, general layers. When not using transfer learning and instead training from scratch (starting with random model weights), discriminative learning rates were not used. Through hyperparameter tuning of all the target phenotypes, the learning rates were found to differ across phenotypes, but were the same across the 2D and 3D models.

When benefitted from transfer learning, the final fully connected layer with softmax output is changed to have the same number of classes as our target classification task, which is 2 for all our prediction settings. Then, we assigned a starting learning rate to our fully connected layer and divided the learning rates for each block before the current block to have a learning rate that is half of the learning rate of the current block. In this manner, the learning rates were inversely decreased as a function of the depth of the blocks. All our transfer learning techniques utilized discriminative learning rates with consistency. When training from scratch, one learning rate was used for all layers.

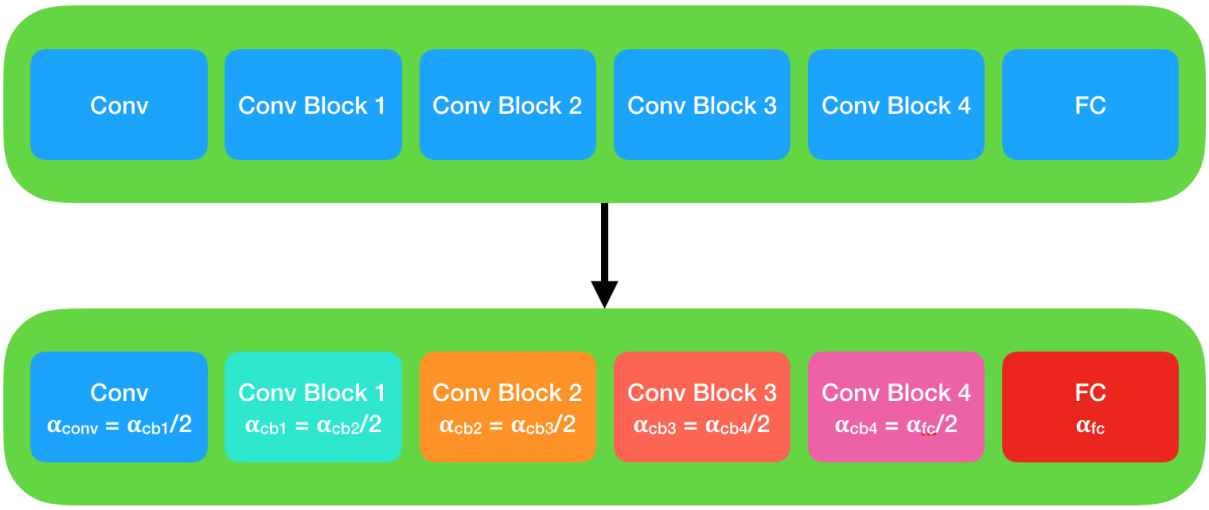

Figure 1: Illustration of discriminative learning rates.

When training from scratch (above), the learning rate is the same across all layers and convolutional blocks for a representative ResNet-18 model. When using discriminative learning rates (below), the highest layer starting from the fully connected layer has the largest learning rate. The convolutional block prior to the last fully connected layer has a learning rate that is half of the fully connected layer. Similarly, earlier convolutional blocks have decreasing learning rates as well, with the first convolutional block having the smallest learning rate.

Table 2: Learning rates used for discriminative learning rates and training from scratch differed based on the target phenotype.

There were two sets of learning rates that were used to train the DNNs: one for training the models from scratch (first column) and another for transfer learning to fine-tune using discriminative learning rates (second column). Hyperparameter tuning was done on each of the phenotypes (rows). When training from scratch, the learning rate was kept constant across the layers of the model (for example, 0.0001 for the sex phenotype). However, when using transfer learning and using discriminative learning rates, in order to preserve the the most general knowledge in the lower layers and modify the higher layers more in comparison (cf. above),

groups of convolutional layers organized in blocks in the ResNet models (cf. below) had different learning rates. The starting learning rate indicates the learning rate of the last fully connected layer (for example, 0.001 for the sex phenotype). From there, the block before the fully connected layer would have a learning rate that is half of that of the fully connected layer. Similarly, the earlier block would have a learning rate that is half of the block after it and so forth. This process ultimately results in the learning rates of earlier layers in the model to have smaller learning rates compared to the later layers.

Phenotype	**Learning rate**	**Starting learning rate (transfer learning using discriminative learning rates)**
Sex	0.0001	0.001
Age	0.0001	0.001
Number in household	0.001	0.01
Household income	0.001	0.01
Fluid intelligence	0.001	0.01
Past smoking	0.001	0.01

3.7. Deep learning models

All of the deep learning analyses were done using both transfer learning as well as training from scratch in order to observe differences (Figure 2). Analysis scenarios with training from scratch

did not use the pretrained models. The DNNs used were of the residual nets (ResNet) family. Compared to traditional DNNs that stack layers on top of each other, ResNets make use of a residual mapping by introducing skipped connections between layers to overcome the vanishing gradient problem for deeper networks. ResNets with deeper layers have been shown to be more easily trainable than traditional DNNs with state-of-the-art results on benchmark tests (K. He et al., 2016).

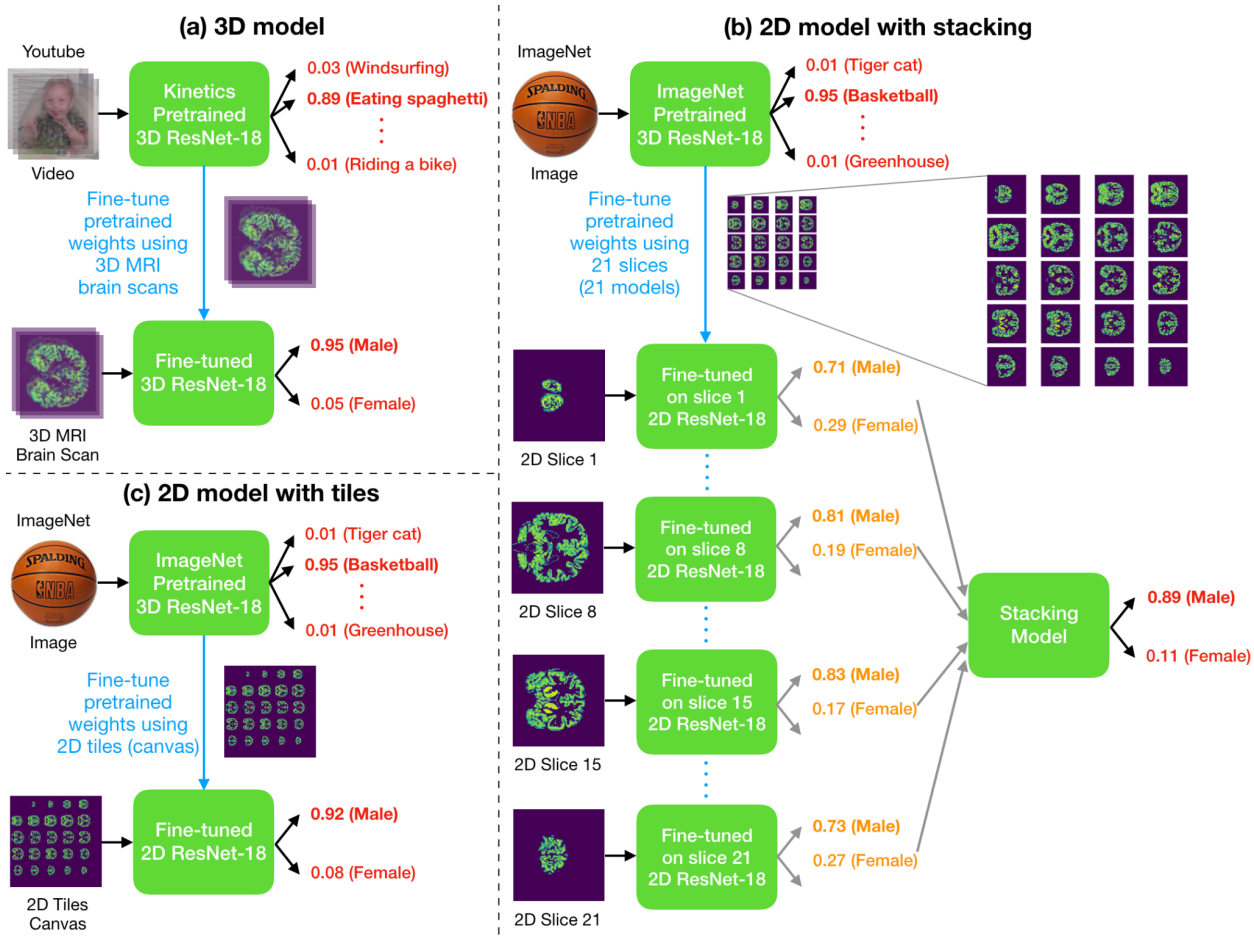

Figure 2: Workflow diagrams for the three distinct types of transfer learning approaches. *We have conducted 3 different types of pipelines to examine the potential of transfer learning in brain-imaging (cf. below). The 3D analyses used a 3D DNN that was fine-tuned on brain images*

after pretraining on 240,000 Youtube videos from the Kinetics dataset, whereas the 2D analyses used a 2D DNN pretrained on 1,280,000 images from the ImageNet dataset. The DNNs were fine-tuned accordingly (cf. below) for each respective workflow. The 3D DNN (a) took in as input full 3D structural brain scans. In contrast, the 2 different 2D analyses had different approaches with (b) having a stacking approach across 21 separate slices and (c) using a canvas of 25 tiles as a single input.

(a) 3D model: The 3D analysis used a 3D ResNet-18 model pretrained on 240,000 Youtube videos from the Kinetics dataset (https://pytorch.org/vision/0.8/models.html#torchvision.models.video.r3d_18). The 3D ResNet-18 model has a total of 33,371,472 model parameters. This neural network architecture is composed of 18 convolution layers with varying layer widths that are broken up in 4 convolution layer blocks. The last fully connected layer uses cross entropy loss. The input data used were the full 3D structural MRI brain scans, with the frames being passed into the model on the sagittal plane in the anterior-posterior direction. When trained on the Kinetics base task, each individual frame from the videos were clipped to a dimension of 112x112. As such, the input data is recommended to have frames of dimension 112x112 for the best performance. The original pretrained model also scaled the video pixel values to a range from 0 to 1 and centered them to 0 mean and scaled the variance to 1 using the mean and standard deviation values derived from the Kinetics dataset (Tran et al., 2018). Accordingly, the 3D MRI scans were resampled to a dimension of 112x134x112 using Nilearn, with the first and third axes being used for the individual frames and the second axis serving as the number of frames or depth in the anterior-posterior direction. Each scan was scaled to have values ranging from 0 to 1. When using transfer learning, the data were standardized using the mean and standard deviation values that are provided by Pytorch based on their pretraining of the 3D ResNet-18 model on

Kinetics. The models were then trained with 3 different random parameter initializations and the out-of-sample prediction accuracy was calculated for each.

(b) 2D model with stacking: The first 2D analysis used a 2D ResNet-18 model pretrained on 1,280,000 images from the ImageNet reference dataset (https://pytorch.org/vision/0.8/models.html#torchvision.models.resnet18). Our model was trained on the 2012 version of the Imagenet classification task which contains 1,280,000 images from 1000 classes. The 2D ResNet-18 model has 11,689,512 total parameters. It is composed of 18 convolution layers with varying layer widths that are broken up into 4 convolution layer blocks. The last fully connected layer uses cross entropy loss. The input data used were 21 different 2D slices of a whole-brain scan along the sagittal plane in the anterior-posterior direction. The slices were selected based on taking the middle slice in the sagittal direction and the respective 10 slices from each side of the middle slice, totaling in 21 slices. The 2D ResNet-18 model also had specific settings used for its training: each image used to train the pretrained model was cropped to a dimension of 224x224, the data were scaled to have a range of 0 to 1 and standardized similarly to (a) (K. He et al., 2016). Akin to the 3D analysis, the scans were resampled, this time to a dimension of 224x268x224, with the first and third axes being used for each individual slice, while the second axis serving as the depth in the anterior-posterior position from which each of the 21 slices were picked. The values were again scaled and standardized similarly to (a). Each of the 21 slices had a respective model that was trained across 3 different model initializations. In order to aggregate the results from all 21 slices into a single outcome, stacking was used. Stacking is an ensemble model averaging technique (Hastie et al., 2009) that we have used in our previous research (Karrer et al., 2019). Multiple base models can be used for separate classifications to then blend the results into a top-level, composite model for a final prediction. The goal of stacking is hence to combine the outcome predictions on a given data point or observation of multiple base models by learning a top-level

linear combination between them to further improve the overall prediction accuracy (Hastie et al., 2009). For our stacking analysis, instead of using a single slice, we were able to combine and estimate the optimal weights from each of the 21 models per slice. The log odds derived from the validation set from the 21 different DNNs per slice, which served as the base models, were used to train a logistic regression model serving as the top-level stacking model. The C value of the logistic regression that controls the regularization strength was kept to the default value of 1. To get the final out-of-sample prediction accuracy, the test set softmax probabilities from each of the 21 base models were used to get the final predictions from using the logistic regression.

(c) 2D model with tiles: The second 2D analysis used the same a 2D ResNet-18 model as (b) that was pretrained on 1,280,000 images from the ImageNet dataset (https://pytorch.org/vision/0.8/models.html#torchvision.models.resnet18). The input data was a canvas of dimension 1000x1000 that was filled with 25 2D slices (tiles) from the sagittal plane as the data used to train, validate and test the DNNs. The tiles were selected based on the dimension of each slice as compared to the dimension of the full 1000x1000 canvas. Filling the canvas from the top left to the bottom right in a row by row manner, the tiles were fitted on the full canvas on a 5x5 grid, resulting in 25 total tiles selected in an incremental step manner to cover all of the slice indexes. Although the input data in this case is quite different than the full 3D scan and single 2D slices from (a) and (b), this tiling method manages to include more information on a single image than (b) while being more computationally efficient since we no longer needed to train one model per slice. Similar to (b), the values were scaled to have a range of 0 to 1 and standardized when using transfer learning. The canvases were used as the constructed conglomerate input data. The models were then trained with 3 different random parameter initializations with each having an out-of-sample prediction accuracy.

4. Results

We leveraged transfer learning using real-world videos and images using the Kinetics and ImageNet datasets for all of our analyses. The massive amounts of data used to pretrain the DNNs provided a robust set of features to further fine-tune to our brain imaging target tasks. We also trained a Linear SVC as the linear baseline to compare against the more complex, non-linear DNNs and examine if there are further non-linear patterns that can be extracted from the data. Linear models have long been a standard used for data analysis due to computational limitations (Bzdok, 2017; Bzdok & Ioannidis, 2019; Efron & Hastie, 2016). Based on previous research, in a variety of common data-analysis scenarios, linear models can perform similarly well as deep learning models in classification tasks utilizing the UKBB dataset in common analysis scenarios (Schulz et al., 2020). With the ~40,000 participant release of the UKBB Imaging resource, we could train more complex DNNs in various ways to further examine the levels of prediction capabilities across a range of phenotype targets and sample sizes.

4.1. Sex classification

Our first analysis focused on sex classification, which is a binary classification task between males and females (Figure 3, Table 3). We chose sex classification to start our investigation due to several reasons. To start, it is the simplest of our examined phenotypes which enabled us to evaluate our linear and non-linear models in a controlled manner. Previous research also showed sex to be the most predictable among other phenotypes (Schulz et al., 2020). More generally, as we examined sex difference in neuroimaging, starting our investigation with sex classification was apt. All our analyses included sample sizes of 5,000 and 40,000 subjects in order to identify any present scaling behavior.

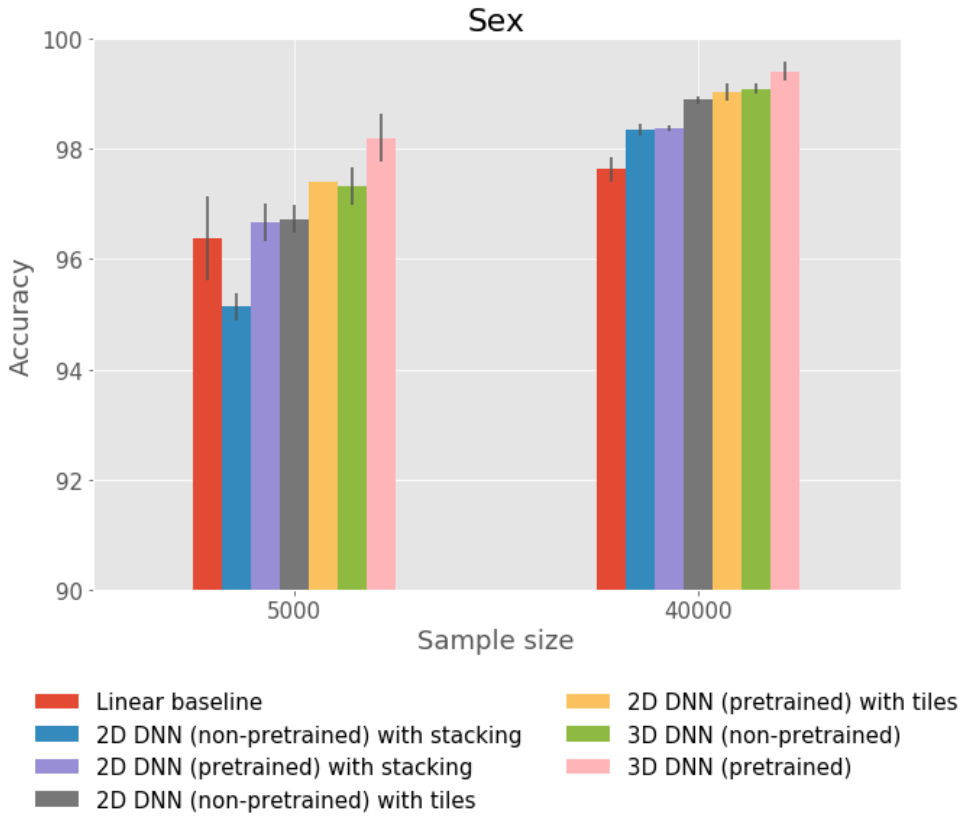

Figure 3: Transfer learning improves sex classification accuracy from brain scans in three different deep learning architectures.

The first analysis was based on sex classification which served as an ingredient for later analyses. The 3D DNNs were pretrained using 240,000 Youtube videos from the Kinetics dataset and the 2D DNNs were pretrained on 1,280,000 images from the ImageNet dataset. The x-axis shows two sample sizes to observe scaling behaviour and the y-axis shows the test set out-of-sample prediction accuracy across the different models. The linear baseline error bars are the standard deviations around the mean across nested cross validations. The DNN error bars are standard deviations around the mean of out-of-sample prediction accuracies across 3 model initializations. DNNs pretrained on both Kinetics (3D) and ImageNet (2D) outperformed DNNs that were trained from scratch. The improved performance from transfer learning was greater on the smaller sample size.

Table 3: Results from sex classification.

	Linear baseline	**2D DNN (non-pretrained) with stacking**	**2D DNN (pretrained) with stacking**	**2D DNN (non-pretrained) with tiles**	**2D DNN (pretrained) with tiles**	**3D DNN (non-pretrained)**	**3D DNN (pretrained)**
Sex (smaller sample size)	96.38% ± 0.77%	95.13% ± 0.25%	96.67% ± 0.34%	96.73% ± 0.25%	97.40% ± 0.00%	97.33% ± 0.34%	**98.20% ± 0.43%**
Sex (larger sample size)	97.63% ± 0.23%	98.35% ± 0.10%	98.37% ± 0.06%	98.89% ± 0.06%	99.03% ± 0.16%	99.10% ± 0.08%	**99.41% ± 0.17%**

The 3D pretrained DNNs with transfer learning (98.20% ± 0.43% with the smaller sample size and 99.40% ± 0.17% with the larger sample size) outperformed all candidate models, consistently across both sample-size scenarios. We posit a few reasons for this. The Kinetics dataset with its vast array of real world videos created a rich set of features that were aided by further fine-tuning (cf. Methods) as the pretrained version of the 3D DNN outperformed its non-pretrained (trained from scratch) counterpart. Additionally, when using the 3D DNNs, the full 3D information from the MRI brain scan was treated as a series of consecutive image frames akin to a video, which preserved all of the information. In contrast, the 2D DNNs only used parts of the full scan. We observed that the non-pretrained 3D DNN (97.33% ± 0.34% with the smaller sample size and 99.10% ± 0.08% with the larger sample size) outperformed all other non-pretrained 2D DNNs as well as the linear baseline. Using the full MRI scan in conjunction with the rich features acquired from the Kinetics dataset resulted in the 3D pretrained DNNs having the best out-of-sample prediction accuracies. Although the 2D DNNs underperformed the 3D DNNs, the pretrained versions of both the stacking and tiles DNNs performed better than the non-pretrained counterparts. This constellation of findings is similar to that of the 3D DNNs, except in the case of the 2D DNNs, the ImageNet dataset was used to create the pretrained weights in the 2D regime. The pretrained 2D DNNs with the 25 tiles canvas (97.40% ± 0.00% with the smaller sampler size and 99.03% ± 0.16% with the larger sample size) performed better than the stacking models with 21 slices (96.67% ± 0.34% with the smaller sample size and 98.37% ± 0.06% with the larger sample size). There could be a few reasons for this: the tiles canvas had the information of 4 additional slices and the architecture provided an end to end deep learning solution instead of an additional stacking procedure. As a whole, transfer learning outperformed training from scratch for all 3 types of DNNs in our sex classification analyses.

In fact, for the smaller 5,000 sample size, the linear baseline (96.38% ± 0.77%) outperformed the non-pretrained 2D DNN with stacking (95.13% ± 0.25%). When using transfer learning with

the 2D DNN with stacking (96.67% ± 0.34%), the performance was on par with the linear baseline which is important to note due to several reasons. For the smaller sample size, using transfer learning in this instance was required to perform at a similar level to the linear baseline for this particular model which could be due to a lack of training observations for the DNN. In addition, the non-pretrained 2D DNN with tiles (96.73% ± 0.25%), with more information per canvas, still only performed similarly to the linear baseline. When using the 2D DNNs with the smaller sample size, we were only able to outperform the linear baseline when using the 25 tiles canvas in conjunction with transfer learning. In the 2D and smaller sample size regime, transfer learning was required to outperform the linear baseline for sex classification.

4.2. Classifying more complex phenotypes

Following sex classification, we directed attention to more complex phenotypes. We divided the phenotypes into two groups by ascending difficulty, and, in each instance, models were fitted by sex. The first group included i) age, ii) number of people in household and iii) household income (Figure 4, Table 4). Although more complex than just sex classification, these phenotypes have shown to be predictable in previous research (Kiesow et al., 2020; Schulz et al., 2020). Along with sex, age is also known to be one of the most salient sources of variability in neuroimaging data, which is hence among the phenotypes that are easiest to predict from brain scans (Ritchie et al., 2018). The second group of phenotypes included i) fluid intelligence and ii) past smoking (Figure 5, Table 5). All of the complex phenotype measures were dichotomized to maximize class balance and enable binary classification. Due to the complex phenotypes being split up by sex, the larger sample size in these instances involved brain images from 20,000 subjects while the smaller sample size remained the same as sex classification with 5,000 subjects.

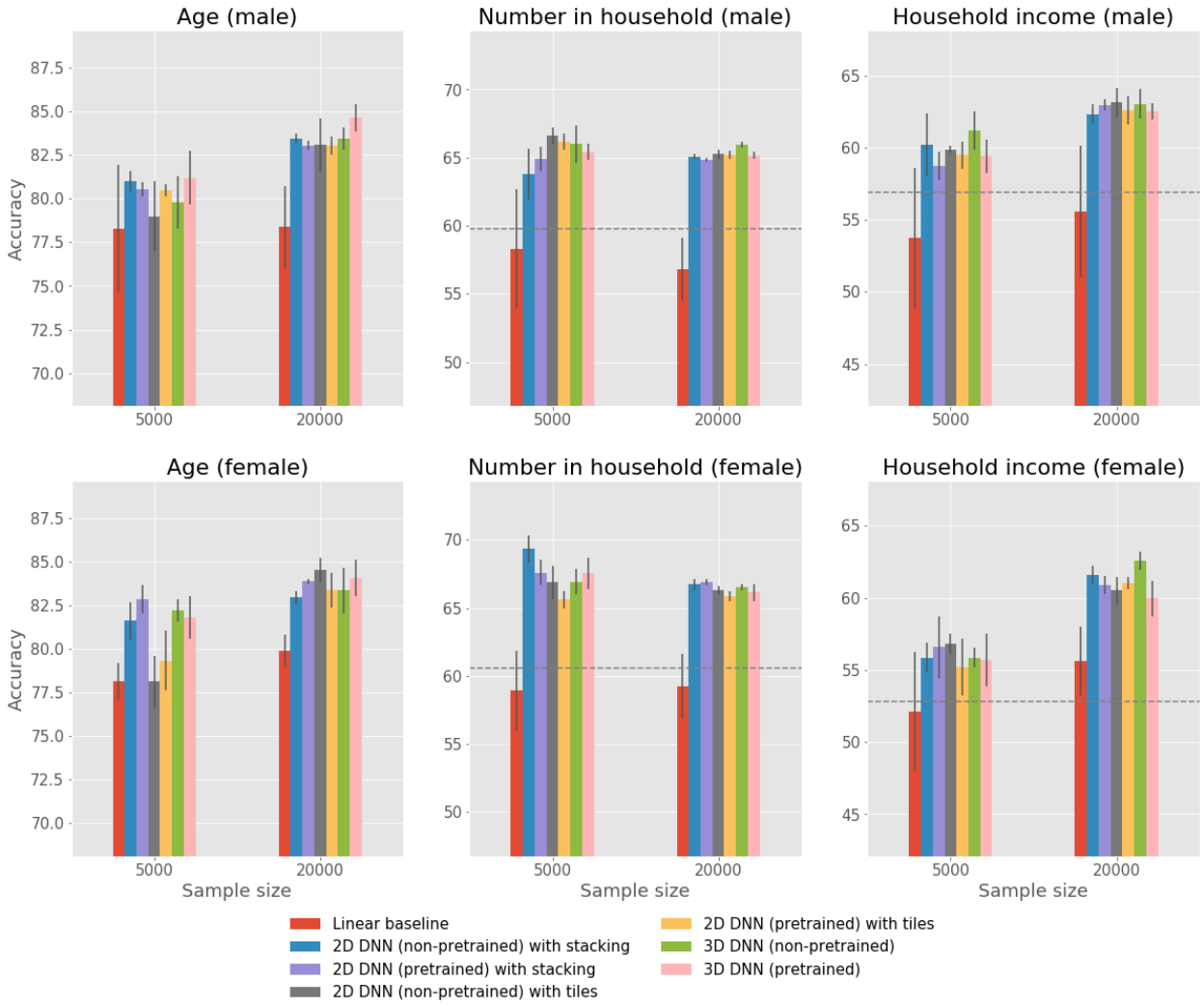

Figure 4: Age exhibited benefits from transfer learning while number in household and household income showcased clear classification differences between males and females.

The first group of complex phenotypes were examined to observe any effects regarding how transfer learning functions on increasing levels of prediction difficulty. The figure shows out-of-sample prediction accuracies for the age, number in household and household income phenotypes split by sex. The x-axis shows two sample sizes to observe scaling behaviour and the y-axis shows the test set out-of-sample prediction accuracy across the different models. The linear baseline error bars are the standard deviations around the mean across nested cross

validations. The DNN error bars are standard deviations around the mean of out-of-sample prediction accuracies across 3 model initializations. The horizontal dashed line (grey), when present, is the chance accuracy. Among the first group of complex phenotypes, age is the most predictable phenotype and it also shows effectiveness with transfer learning.

Table 4: Results from age, number in household and household income classifications. Each phenotype was split by sex.

	Linear baseline	**2D DNN (non-pretrained) with stacking**	**2D DNN (pretrained) with stacking**	**2D DNN (non-pretrained) with tiles**	**2D DNN (pretrained) with tiles**	**3D DNN (non-pretrained)**	**3D DNN (pretrained)**
Age (male, smaller sample size)	78.28% ± 3.63%	81.00% ± 0.57%	80.53% ± 0.41%	79.00% ± 2.01%	80.47% ± 0.34%	79.80% ± 1.50%	**81.20% ± 1.56%**
Age (male, larger sample size)	78.39% ± 2.35%	83.43% ± 0.28%	83.05% ± 0.25%	83.07% ± 1.52%	83.05% ± 0.53%	83.45% ± 0.65%	**84.64% ± 0.78%**
Age (female, smaller sample size)	78.14% ± 1.07%	81.60% ± 1.07%	**82.87% ± 0.81%**	78.13% ± 1.48%	79.33% ± 1.73%	82.20% ± 0.65%	81.80% ± 1.23%

Age (female, larger sample size)	79.90% ± 0.92%	82.95% ± 0.37%	83.88% ± 0.14%	**84.53% ± 0.67%**	83.36% ± 0.99%	83.34% ± 1.31%	84.06% ± 1.02%
Number in household (male, smaller sample size)	58.30% ± 4.35%	63.80% ± 1.88%	64.93% ± 0.90%	**66.60% ± 0.59%**	66.20% ± 0.59%	66.00% ± 1.40%	65.40% ± 0.59%
Number in household (male, larger sample size)	56.80% ± 2.31%	65.08% ± 0.18%	64.86% ± 0.16%	65.26% ± 0.35%	65.20% ± 0.27%	**65.95% ± 0.23%**	65.15% ± 0.26%
Number in household (female, smaller sample size)	58.92% ± 2.94%	**69.33% ± 1.00%**	67.60% ± 0.91%	66.87% ± 1.20%	65.60% ± 0.65%	66.93% ± 0.90%	67.53% ± 1.15%

Number in household (female, larger sample size)	59.22% ± 2.39%	66.73% ± 0.42%	**66.93% ± 0.22%**	66.32% ± 0.30%	65.87% ± 0.36%	66.54% ± 0.25%	66.13% ± 0.61%
Household income (male, smaller sample size)	53.72% ± 4.86%	60.20% ± 2.20%	58.73% ± 1.00%	59.87% ± 0.25%	59.47% ± 0.96%	**61.20% ± 1.34%**	59.40% ± 1.18%
Household income (male, larger sample size)	55.58% ± 4.55%	62.32% ± 0.65%	62.95% ± 0.39%	**63.11% ± 1.03%**	62.61% ± 0.98%	63.03% ± 1.04%	62.51% ± 0.54%
Household income (female,	52.12% ± 4.11%	55.87% ± 1.04%	56.60% ± 2.14%	**56.80% ± 0.71%**	55.20% ± 1.96%	55.87% ± 0.66%	55.67% ± 1.82%

smaller sample size)							
Household income (female, larger sample size)	55.63% ± 2.38%	61.58% ± 0.62%	60.91% ± 0.64%	60.52% ± 0.95%	61.05% ± 0.41%	**62.56% ± 0.64%**	59.96% ± 1.22%

Whereas sex classification demonstrated clear benefits of transfer learning as well as the 3D DNNs performing the best, we did not observe the same constellation of findings for the more complex phenotypes. Among the target phenotypes of age, number in household and household income, age was by far the most predictable phenotype, for both males and females. The DNNs outperformed the linear baselines especially with the larger sample size, but there was not a single model that outperformed all other models consistently. For age among males, the pretrained 3D DNNs (81.20% ± 1.56% with the smaller sample size and 84.64% ± 0.78% with the larger sample size) achieved the best prediction accuracies across sample sizes, but among females for the smaller sample size, the pretrained 2D DNN with stacking (82.87% ± 0.81%) performed the best and as for the larger sample size, the non-pretrained 2D DNN with tiles (84.53% ± 0.67%) and the pretrained 3D DNN (84.06% ± 1.02%) had similar results. Although there was not a single best performing model, transfer learning did benefit classification for the age phenotype for males and females. Furthermore, the effectiveness of transfer learning in this setting was encouraging for more complex phenotype targets.

Compared to age, number in household and household income exhibited a greater difference in prediction accuracy when split by sex. In terms of number in household, the female splits exhibited higher prediction accuracies across sample sizes, with the best performing models being the non-pretrained 2D DNN with stacking for the smaller sample size (69.33% ± 1.00%) and the pretrained 2D DNN with stacking for the larger sample size (66.93% ± 0.22%). In comparison to males, the best performing model was the non-pretrained 2D DNN with tiles (66.60% ± 0.59%) for the smaller sample size and the non-pretrained 3D DNN (65.95% ± 0.22%) for the larger sample size. However, using the larger sample size did not improve the results as the highest prediction accuracies were from the smaller sample size. In contrast, household income showed the opposite effect where the results from males exhibited higher

out-of-sample prediction accuracies across sample sizes, with the non-pretrained 3D DNN (61.20% ± 1.34%) having the best results for the smaller sample size and the non-pretrained 2D DNN with tiles (63.11% ± 1.03%) performing best on the larger sample size compared to females where the pretrained 2D DNN with stacking (56.60% ± 2.14%) did best on the smaller sample size and the non-pretrained 3D DNN (62.56% ± 0.64%) had the best results on the larger sample size. In comparison to the age phenotype that showcased beneficial usage of transfer learning, the number in household and household income phenotypes did not exhibit a similarly equivocal set of findings.

We then examined further complex phenotypes: fluid intelligence and past smoking (Figure 5, Table 5). These phenotypes were the most challenging classification tasks confronted in this work. Similar to the number in household and household income phenotypes (cf. above), there was no pattern of transfer learning greatly outperforming models that were trained from scratch as the best performing models varied across both models and sample sizes (66.60% ± 0.59% with non-pretrained 2D DNN with tiles for males in number in household in the smaller sample size, 69.33% ± 1.00% with non-pretrained 2D DNN with stacking for females in number in household in the smaller sample size, 63.11% ± 1.03% with non-pretrained 2D DNN with tiles for males in household income in the larger sample size, 62.56% ± 0.64% with non-pretrained 3D DNN for females in household income in the larger sample size, 56.24% ± 0.34% with non-pretrained 2D DNN with stacking for males in fluid intelligence in the larger sample size, 54.73% ± 0.75% with pretrained 2D DNN with stacking for females in fluid intelligence in the larger sample size, 59.27% ± 0.93% with non-pretrained 3D DNN for males in past smoking in the smaller sample size, 53.87% ± 1.32% with non-pretrained 2D DNN with tiles for females in past smoking in the smaller sample size). Whereas sex classification displayed transfer learning efficacy and a singular best performing model, the results from the more complex phenotypes were much more mixed.

Both across males and females, fluid intelligence showed the poorest predictability in both sample-size scenarios, with females having slightly higher prediction accuracies with the pretrained 2D DNN with stacking performing the best on both the smaller (52.73% ± 0.19%) and larger (54.73% ± 0.75%) sample sizes. In comparison, although the past smoking phenotype showed a similar lack of predictability in females with all results being close to chance, there appeared to be a distinct increase in prediction accuracy among males, with the best performing models being the non-pretrained 3D DNN for the smaller sample size (59.27% ± 0.93%) and the non-pretrained 2D DNN with stacking for the larger sample size (59.08% ± 0.64%). Of note, all examined DNN architectures showed overlapping results, which became even more similar using the larger sample size when classifying past smoking in males.

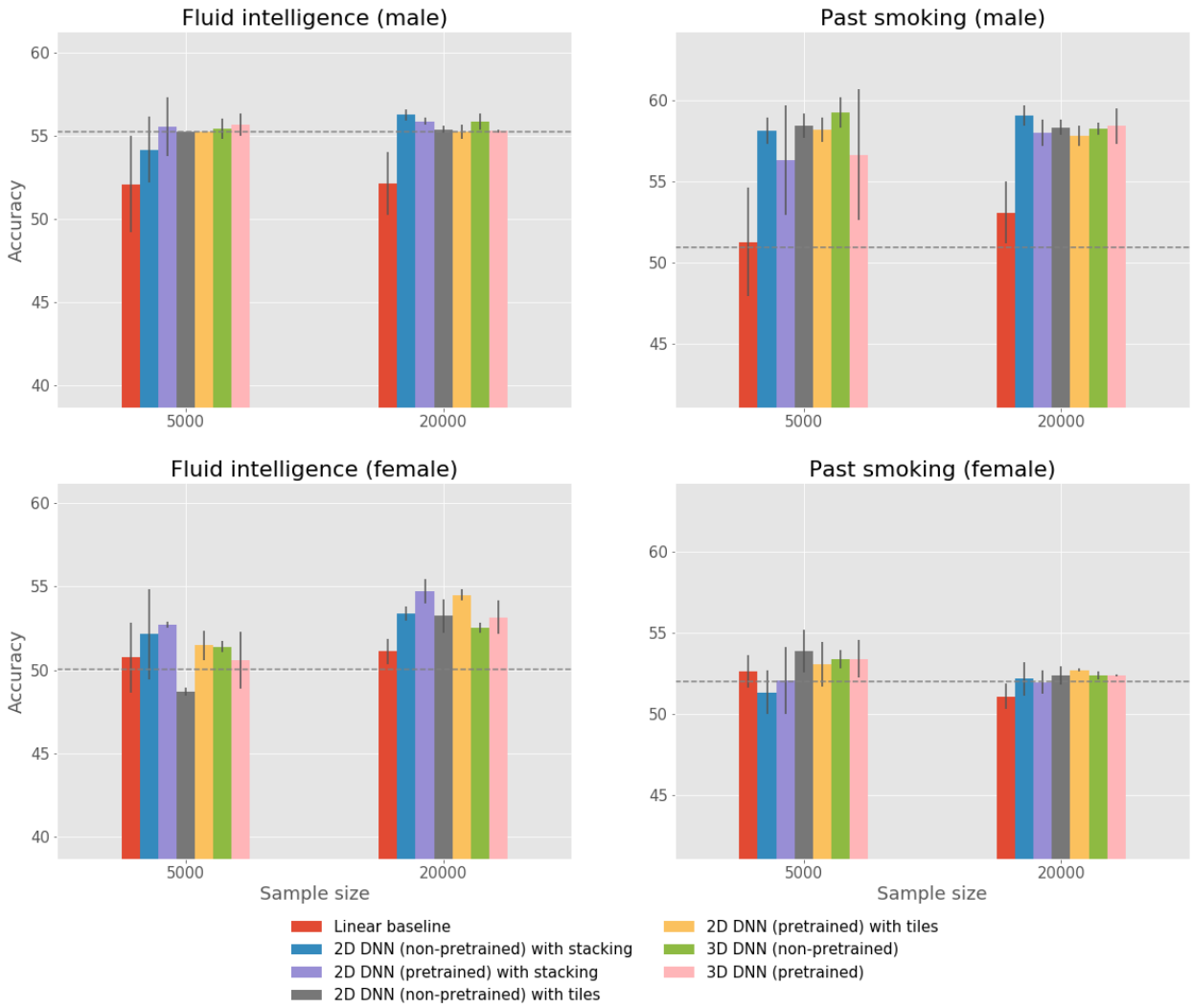

Figure 5: Fluid intelligence and past smoking showed different out-of-sample prediction accuracies in males and females, despite the difficult prediction tasks.

The group of complex phenotypes provided higher levels of prediction difficulty. The figure shows prediction accuracies for the fluid intelligence and past smoking phenotypes split by sex. The x-axis shows two sample sizes to observe scaling behaviour and the y-axis shows the test set out-of-sample prediction accuracy across the different models. The linear baseline error bars are the standard deviations around the mean across nested cross validations. The DNN error bars are standard deviations around the mean of out-of-sample prediction accuracies across 3 model initializations. The horizontal dashed line (grey), when present, is the chance accuracy.

Fluid intelligence and past smoking were the two least predictable phenotypes both in terms of lack of consistency using transfer learning as well as overall out-of-sample prediction accuracy, but there were distinct differences in predictability for males and females.

Table 5: Results from fluid intelligence and past smoking classifications (each phenotype was split by sex).

	Linear baseline	**2D DNN (non-pretrained) with stacking**	**2D DNN (pretrained) with stacking**	**2D DNN (non-pretrained) with tiles**	**2D DNN (pretrained) with tiles**	**3D DNN (non-pretrained)**	**3D DNN (pretrained)**
Fluid intelligence (male, smaller sample size)	52.08% ± 2.90%	54.13% ± 1.98%	55.53% ± 1.79%	55.20% ± 0.00%	55.20% ± 0.00%	55.40% ± 0.59%	**55.67% ± 0.66%**
Fluid intelligence (male, larger sample size)	52.13% ± 1.89%	**56.24% ± 0.34%**	55.84% ± 0.21%	55.37% ± 0.22%	55.22% ± 0.40%	55.84% ± 0.47%	55.26% ± 0.10%
Fluid intelligence	50.74% ± 2.11%	52.13% ± 2.69%	**52.73% ± 0.19%**	48.67% ± 0.25%	51.47% ± 0.90%	51.40% ± 0.33%	50.60% ± 1.70%

(female, smaller sample size)							
Fluid intelligence (female, larger sample size)	51.11% ± 0.76%	53.38% ± 0.40%	**54.73% ± 0.75%**	53.23% ± 1.03%	54.50% ± 0.32%	52.53% ± 0.30%	53.17% ± 1.02%
Past smoking (male, smaller sample size)	51.28% ± 3.34%	58.13% ± 0.82%	56.33% ± 3.40%	58.47% ± 0.77%	58.20% ± 0.75%	**59.27% ± 0.93%**	56.67% ± 4.05%
Past smoking (male, larger sample size)	53.10% ± 1.89%	**59.08% ± 0.64%**	58.00% ± 0.83%	58.36% ± 0.44%	57.82% ± 0.64%	58.25% ± 0.38%	58.45% ± 1.10%

Past smoking (female, smaller sample size)	52.62% ± 0.98%	51.33% ± 1.32%	52.07% ± 2.08%	**53.87% ± 1.32%**	53.07% ± 1.36%	53.40% ± 0.57%	53.40% ± 1.14%
Past smoking (female, larger sample size)	51.07% ± 0.77%	52.16% ± 1.04%	51.96% ± 0.72%	52.37% ± 0.56%	**52.70% ± 0.09%**	52.37% ± 0.22%	52.37% ± 0.07%

It is also important to note the difference between the linear baseline and the deep learning models across phenotype predictions. Among the simpler phenotypes such as sex and age, although the DNNs mostly outperformed the linear baseline, the results were relatively closer in an inverse manner to the complexity of the phenotype. For example, with sex being the simplest to predict of our studied phenotypes, the linear baseline results are closer to the deep learning results. A more complex phenotype than sex, age has a greater disparity of out-of-sample prediction accuracy between the linear baseline and DNNs. For the remaining complex phenotypes (number in household, household income, fluid intelligence and past smoking), the differences in out-of-sample prediction accuracy between the linear baseline and the DNNs are even greater when prediction above chance is possible. This could be due to any exploited non-linearities by the DNNs that are not accessible to the linear model.

5. Discussion

The data scarcity in brain-imaging presents a major challenge to effectively train DNNs in many mission-critical settings. We used emerging transfer learning techniques that learned structured a-priori knowledge (inductive biases) from general purpose datasets: the massive video databases Youtube and the natural images from reference dataset ImageNet. Once trained, the DNN parameters were then fine-tuned to refine the prediction of target tasks with much smaller datasets (Deng et al., 2009; Kay et al., 2017). Although not directly related to brain scans, the vast array of real-world actions depicted by the images and videos can provide the basis for a strong, general feature extractor. By applying transfer learning in combination with the largest biomedical dataset in the world in the UKBB, we show improved DNN predictions out-of-sample.

Sex difference is one of the most salient sources of variability in biology in general and in the human brain in particular (Kiesow et al., 2020; Ritchie et al., 2018). As such, we aimed to apply transfer learning to examine differences exhibited in classification tasks for phenotypes in males and females from structural brain MRI scans. Varying extents of sex difference in the human brain has been shown across many studies. A previous study based on the UKBB, which was the largest study of its kind at the time, has reported sex difference in anatomical measurements (Ritchie et al., 2018). Males were found to have larger volumes and surface area, while females showed thicker cortices in a series of brain regions. The larger brain volumes were found to be greater in some regions associated with emotion and decision making. Despite the differences, there were overlaps in measurement distributions found between males and females. Machine learning has also been previously applied to study specific phenotypes specifically in males or females, such as antisocial behaviour in incarcerated individuals (Anderson et al., 2019). For both datasets, the sex classification accuracy was found to be above 93%, suggesting that in terms of brain imaging data, the male and female brains are differentiable across datasets.

Brain volume has likewise been found to be related to male and female children with early age traumatic life events such as death of a loved one, violence and accidents (Badura-Brack et al., 2020). There was no main effect in regional volumes observed on the whole sample. However, when broken out by sex, the authors reported significant differences: in the high trauma group, girls exhibited greater volumes in the hippocampal and parahippocampal regions than boys, with follow-up analyses showing increasing volumes for girls and decreasing volumes for boys.

Fluid intelligence is one of the phenotypes we examined in this study and it has also been studied for its development in adolescents (Taylor et al., 2020). When examining the sample as a whole, these authors found that older participants had stronger theta activity around the calcarine fissure and in the cerebellar cortices compared to younger participants. More specifically, when examining the sexes separately, although the mean performance for males and females were found to be similar, males exhibited increased theta activity associated with increasing age and faster reaction times, whereas females had decreased theta activity with increasing age and better task accuracy. The authors also point out similarities of their findings in conjunction with the Parieto-Frontal Integration Theory of intelligence (Jung & Haier, 2007), noting that sex difference may play a significant role in the abstraction of fluid intelligence and reasoning. The largest study of its kind examined phenotypes centered around social stimulation (Kiesow et al., 2020). Complementary to Ritchie et al., although some degree of overlap between males and females was observed, there were also differences found between the sexes in phenotypes such as the number of people in the household and socioeconomic status, both of which we examine in this study as well. Notably, the dissimilarities between males and females in social interactions may hold significance in how males and females have survived and evolved in social settings.

Our objective was twofold. First, we applied transfer learning to more effectively train DNNs of various levels of complexities in the data scarce brain imaging setting and compared the results against DNNs trained from scratch as well as a linear baseline. Second, we further investigated the effects of sex differences on machine learning pipelines, which has been suggested by previous studies to be one of the key sources of variability in brain imaging across datasets and analyses, on phenotypes of varying complexities (Schulz et al., 2020). Data scarcity in brain imaging poses challenges to effectively exploit DNNs that require a large amount of data to train robust features. In order to facilitate training of DNNs on brain imaging data, we applied transfer learning using the Kinetics and ImageNet datasets, which are widely used to pretrain DNNs in various kinds of tasks including ones in medical imaging, to build separate models for males and females. More intuitively, we can think of humans experiencing real world events over time to develop our vision which certainly plays a part in a medical health professional's ability to view and assess MRI brain scans.

To harness the learned structured knowledge encoded within the pretrained DNN parameters, transfer learning has been used on many medical classification tasks. Transfer learning using pretrained weights from the Kinetics dataset has been found to improve the diagnosis of appendicitis from abdominal CT scans with an AUC of 0.810 (95% CI 0.725, 0.895) compared to training from scratch with an AUC of 0.724 (95% CI 0.625, 0.823) (Rajpurkar et al., 2020). The authors also note that the effectiveness of transfer learning would likely diminish with more training samples in the target dataset, reiterating more general principles of transfer learning regarding the size of the target dataset (Yosinski et al., 2014). Similarly, ImageNet pretrained DNNs have previously aided classification of Alzheimer's disease using brain MRI scans from the ADNI dataset (Jack et al., 2008; Khan et al., 2019). The DNN was able to achieve state-of-the-art results on both binary classification and a 3-class classification task between Alzheimer's Disease, mild cognitive impairment and normal control. The analysis used a small

training dataset to achieve the results to emphasize the efficacy of transfer learning. For the binary classification, the transfer learning approach showed improvements of 4% and 7% over the previous state of the art and the 3-way classification task using transfer learning had an improved prediction accuracy of 95.19% over 89.1% using random weights. The authors also discuss freezing parts of the DNN, which relates to our usage of discriminative learning rates (cf. Methods) to keep the general acquired knowledge from the pretraining step. There has also been work done on multi-class classification of abdominal ultrasound images using ImageNet pretrained DNNs (Cheng & Malhi, 2017). Comparing the classification results between a trained radiologist and the DNNs, the models showcased a higher classification accuracy of 77.9% on the test set compared to the radiologist's 71.7%. The authors also noted that transfer learning could effectively be used to train classification models for abdominal ultrasound images.

Transfer learning has also shown some mixed results in our analyses. A previous study focusing on retinal fundus images to detect diabetic retinopathy and chest x-rays to detect 5 different pathologies found that on the retina task, transfer learning performed comparably to training from scratch, whereas the chest x-ray tasks had mixed results based on pathologies (Raghu et al., 2019). The lack of consistent improvements using transfer learning on certain tasks may signal the level of difficulty of a particular target task could potentially aid or hinder transfer learning. Additionally, when trained on a smaller subset of the retina dataset, transfer learning performed better than training from scratch with a prediction accuracy of 94.6% compared to 92.2%, with the larger models showing the best results compared to the smaller models. The authors propose the reason for the larger models performing better with transfer learning could be due to over-parameterization as well as feature reuse. Over-parameterization would occur if the number of parameters would be quite large compared to the training dataset, which could potentially introduce "memorization" of the data. There is also evidence of feature reuse of more general features learned from ImageNet such as edges and curves. It was found that most

meaningful features were from the lower layers containing the most general features, which is our reason for using discriminative learning rates to train our DNNs. We took the learnings from previous research of transfer learning on medical datasets and applied them to our brain imaging data.

In our analyses, sex classification, which is the simplest of our 6 target phenotypes, showed obvious benefits from using transfer learning across sample sizes. For both the smaller and larger sample sizes, the pretrained 3D DNNs using transfer learning outperformed all other models. The pretrained version of each model also outperformed its non-pretrained counterpart for both sample sizes with a higher rate of improvement for the smaller sample size, a detail consistent with other studies. Based on the results from sex classification, transfer learning appears to be more effective on smaller datasets compared to larger datasets on two facets: transfer learning was required to outperform the linear baseline when using the 2D DNNs and relatively, comparing the non-pretrained and pretrained version of each model, we observed a more significant prediction improvement for the smaller sample size (1.53 p.p., 0.67 p.p., 0.87 p.p. improvements respectively for the 2D DNN with stacking, 2D DNN with tiles and 3D DNN for the smaller sample size compared to 0.02 p.p., 0.14 p.p., 0.31 p.p. for the larger sample size). Although a larger training set would be expected to produce higher prediction accuracies, the diminishing efficacy of transfer learning for the larger sample size can be explained by revisiting the original reasoning for using transfer learning: to train models more effectively using smaller sample sizes. It is possible that the current full UKBB dataset could have enough observations to effectively train DNNs for sex classification. The encouraging results of transfer learning on sex classification could signal that transfer learning could improve prediction accuracy on brain imaging data. In order to further investigate this notion, we applied transfer learning to additional phenotypes broken up by sex.

The more complex phenotypes in our analyses included age, number in household, household income, fluid intelligence and past smoking. Our results were mixed with regards to out-of-sample prediction accuracies on the complex phenotypes. Age, which is the second simplest phenotype we examined, showed some benefits of transfer learning among males across sample sizes and for the smaller sample size for females. Generally, the prediction accuracies for the age phenotype in males and females were overall similar. However, for the rest of the phenotypes (number in household, household income, fluid intelligence and past smoking), we did not find any consistent pattern of efficacy when transfer learning was applied to either sex or sample size, with different models performing the best for different phenotypes. There was no pattern of greater prediction accuracy improvement in the smaller sample size when transfer learning was used nor was there any consistent, best performing model across sample sizes. It is interesting to note as well that these complex phenotypes in general are difficult prediction tasks. In conjunction with the lack of pattern using transfer learning, it is possible that the general features that are extracted from the base dataset may not be as useful in such settings. As Raghu et al. discuss in their work regarding feature reuse, it could be possible that despite preserving the knowledge in the lower layers, for more complicated tasks, it is oftentimes difficult to outperform training from scratch.

Although transfer learning did not exhibit clear benefits when predicting complex phenotypes in terms of improved out-of-sample prediction accuracy, there were differences in the overall results in males and females. For the number in household phenotype, females showed higher out-of-sample prediction accuracies for the smaller and larger sample sizes; a detail that is consistent with previous research showing household sizes were associated with greater brain volume overlap in males (Kiesow et al., 2020), which could mean lower variability. Features with high variability can help a machine learning model exploit predictive patterns in the dataset. Conversely, lower variability could in turn lower predictive power. The household income

phenotype had the opposite effect with males having better out-of-sample prediction accuracies. Again, the results were in line with previous research which found there to be greater overlap in brain volume in association with household income among females which could suggest lower variability (Kiesow et al., 2020) and predictive power. Although difficult classification tasks, the number in household and household income phenotypes showed divergent patterns in males and females, whereas fluid intelligence was found to be the most challenging phenotype. In several meta-analyses examining cognitive skills that could be related to fluid intelligence (Feingold, 1988; Hines, 2010; Janet S. Hyde & Linn, 1988; J. S. Hyde et al., 1990), it was found for tasks and categories potentially related to fluid intelligence such as SAT mathematics, computational skills, math concepts, verbal fluency, perceptual speed, vocabulary and SAT verbal, there was very little to no differences between males and females. Hyde et al. also note that the magnitude of any differences in the measured tasks tend to diminish over time. By the same logic, with the recent UKBB dataset, we could expect decreasing differences in such phenotypes across sexes. The difficulty of the fluid intelligence classification task is also consistent with previous research (Schulz et al., 2020). In comparison to previous analyses carried out by Schulz et al., we observed a similar pattern of level of difficulty for prediction tasks, with sex classification performing the best, age being the second simplest phenotype to predict, number in household and household income being of similar difficulty following age, and fluid intelligence being the least predictable of our set of phenotypes. Additionally, we examined past smoking in the current study where males had greater out of sample prediction accuracies compared to females who had accuracies around chance. There have been other studies showing greater reactivity to smoking in males compared to females (Dumais et al., 2017). However, it has also been noted that smoking differences among males and females have decreased over time (Peters et al., 2014), so future work could include further splits by both sex and age to observe differences. In all of the stated phenotypes, it is obvious there is a drop in predictive performance after sex and age classification. As studied before by Ritchie et al., there

are significant distribution overlaps in brain measurements in males and females which could lead to challenges in predicting certain phenotype targets.

Our overarching goal has been to combine state-of-the-art transfer learning techniques along with the UKBB dataset to enhance training of DNNs in the data-scarce brain imaging setting. Moreover, we divided the dataset by sex for a range of phenotype complexities to observe sex differences and most successfully train our machine learning models. Although we showcased both the success of transfer learning as well as shortcomings in our selected phenotypes, we also observed distinct sex differences in out-of-sample prediction performance in our results. However, the importance of studying sex differences goes beyond raw prediction capabilities. Despite being one of the most conserved differences through evolution, potential gains of acknowledging sex difference in machine learning pipelines has remained neglected in research (Klein et al., 2015). Incorporating sex differences in analyses can potentially help in discovery and innovation in science and engineering (Tannenbaum et al., 2019). Conforming to the status quo of experimental designs often ignores sex differences in datasets. More thorough analyses split up by sex could potentially surface patterns that would otherwise be hidden for some of the greatest challenges facing us in the 21st century including human therapeutics, safer products, reducing AI bias, equalizing stereotypes and reproducibility of algorithms (Tannenbaum et al., 2019). Therefore, it is key for us to understand the minutiae of sex differences to better train machine learning models using transfer learning in order to ultimately affect positive changes.

6. Conclusion

The focus of this study has been to draw upon the state of the art transfer learning techniques to improve brain imaging classification tasks to overcome important challenges imposed by data scarcity. To this end, the integration of transfer learning with brain imaging shows promise in utilizing deep learning to maximize out-of-sample prediction accuracies with more room for improvement in fine-tuning models with more complex phenotypes. Although this study did not focus on clinical tasks, it could be possible to extend models to tasks such as diagnosis detection, risk prediction and automated treatment selection. There are many examples of studies focusing on clinical tasks in the brain imaging setting, but it is paramount that the model pipelines are correctly implemented for specific diseases and tasks due to issues such as incorrect cross-validation and overfitting (Mateos-Pérez et al., 2018). The added nuance of sex differences in datasets could also play a significant role in exploiting existing variabilities in datasets as is the case in autism (Baron-Cohen et al., 2005). However, today's biggest problems such as precision medicine and AI in healthcare exhibit ever increasing complexities. In order to better tackle such challenges, instead of applying singular facets such as larger datasets, certain deep learning models or data that is split by sex, it is more likely that richer solutions composed of various different methods and techniques could be required. Transfer learning is an increasingly active area of research that is focused on more efficient AI and it could be key in imaging neuroscience advancements.

References

Aggarwal, R., Sounderajah, V., Martin, G., Ting, D. S. W., Karthikesalingam, A., King, D., Ashrafian, H., & Darzi, A. (2021). Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digital Medicine*, *4*(1), 65.

Anderson, N. E., Harenski, K. A., Harenski, C. L., Koenigs, M. R., Decety, J., Calhoun, V. D., & Kiehl, K. A. (2019). Machine learning of brain gray matter differentiates sex in a large forensic sample. *Human Brain Mapping*, *40*(5), 1496–1506.

Badura-Brack, A. S., Mills, M. S., Embury, C. M., Khanna, M. M., Klanecky Earl, A., Stephen, J. M., Wang, Y.-P., Calhoun, V. D., & Wilson, T. W. (2020). Hippocampal and parahippocampal volumes vary by sex and traumatic life events in children. *Journal of Psychiatry & Neuroscience: JPN*, *45*(4), 288–297.

Baron-Cohen, S., Knickmeyer, R. C., & Belmonte, M. K. (2005). Sex differences in the brain: implications for explaining autism. *Science*, *310*(5749), 819–823.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Biswas, M., Kuppili, V., Saba, L., Edla, D. R., Suri, H. S., Cuadrado-Godia, E., Laird, J. R., Marinhoe, R. T., Sanches, J. M., Nicolaides, A., & Suri, J. S. (2019). State-of-the-art review on deep learning in medical imaging. *Frontiers in Bioscience* , *24*, 392–426.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209.

Bzdok, D. (2017). Classical Statistics and Statistical Learning in Imaging Neuroscience. *Frontiers in Neuroscience*, *11*. https://doi.org/10.3389/fnins.2017.00543

Bzdok, D., Engemann, D., & Thirion, B. (2020). Inference and Prediction Diverge in Biomedicine. In *Patterns* (Vol. 1, Issue 8, p. 100119).

https://doi.org/10.1016/j.patter.2020.100119

Bzdok, D., & Ioannidis, J. P. A. (2019). Exploration, Inference, and Prediction in Neuroscience and Biomedicine. *Trends in Neurosciences*, *42*(4), 251–262.

Bzdok, D., Varoquaux, G., & Steyerberg, E. W. (2021). Prediction, Not Association, Paves the Road to Precision Medicine. *JAMA Psychiatry* , *78*(2), 127–128.

Cahill, L. (2014). Fundamental sex difference in human brain architecture [Review of *Fundamental sex difference in human brain architecture*]. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(2), 577–578.

Cai, B., Zhang, G., Zhang, A., Hu, W., Stephen, J. M., Wilson, T. W., Calhoun, V. D., & Wang, Y.-P. (2020). A GICA-TVGL framework to study sex differences in resting state fMRI dynamic connectivity. *Journal of Neuroscience Methods*, *332*, 108531.

Cheng, P. M., & Malhi, H. S. (2017). Transfer Learning with Convolutional Neural Networks for Classification of Abdominal Ultrasound Images. *Journal of Digital Imaging*, *30*(2), 234–243.

Chung, Y. S., Calhoun, V., & Stevens, M. C. (2020). Adolescent sex differences in cortico-subcortical functional connectivity during response inhibition. *Cognitive, Affective & Behavioral Neuroscience*, *20*(1), 1–18.

de Lacy, N., McCauley, E., Kutz, J. N., & Calhoun, V. D. (2019). Sex-related differences in intrinsic brain dynamism and their neurocognitive correlates. *NeuroImage*, *202*, 116116.

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Dumais, K. M., Franklin, T. R., Jagannathan, K., Hager, N., Gawrysiak, M., Betts, J., Farmer, S., Guthier, E., Pater, H., Janes, A. C., & Wetherill, R. R. (2017). Multi-site exploration of sex differences in brain reactivity to smoking cues: Consensus across sites and methodologies. *Drug and Alcohol Dependence*, *178*, 469–476.

Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.

Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., & Dehmer, M. (2020). An Introductory Review of Deep Learning for Prediction Models With Big Data. *Frontiers in Artificial Intelligence*, *3*, 4.

Feingold, A. (1988). Cognitive gender differences are disappearing. *The American Psychologist*, *43*(2), 95.

Foster, K. R., Koprowski, R., & Skufca, J. D. (2014). Machine learning, medical diagnosis, and biomedical engineering research - commentary. *Biomedical Engineering Online*, *13*, 94.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science & Business Media.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

He, T., Kong, R., Holmes, A. J., Sabuncu, M. R., Eickhoff, S. B., Bzdok, D., Feng, J., & Yeo, B. T. T. (2018). Is deep learning better than kernel regression for functional connectivity prediction of fluid intelligence? *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 1–4.

Hines, M. (2010). Sex-related variation in human behavior and the brain. *Trends in Cognitive Sciences*, *14*(10), 448–456.

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 328–339.

Hussain, M., Bird, J. J., & Faria, D. R. (2019). A Study on CNN Transfer Learning for Image Classification. *Advances in Computational Intelligence Systems*, 191–202.

Hutchinson, M. L., Antono, E., Gibbons, B. M., Paradiso, S., Ling, J., & Meredig, B. (2017). Overcoming data scarcity with transfer learning. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1711.05099

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: a meta-analysis. *Psychological Bulletin*, *107*(2), 139–155.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*(1), 53.

Iraji, A., Faghiri, A., Fu, Z., Rachakonda, S., Kochunov, P., Belger, A., Ford, J. M., McEwen, S., Mathalon, D. H., Mueller, B. A., Pearlson, G. D., Potkin, S. G., Preda, A., Turner, J. A., van Erp, T. G. M., & Calhoun, V. D. (2021). Multi-Spatial Scale Dynamic Interactions between Functional Sources Reveal Sex-Specific Changes in Schizophrenia. In *Cold Spring Harbor Laboratory* (p. 2021.01.04.425222). https://doi.org/10.1101/2021.01.04.425222

Jack, C. R., Jr, Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L Whitwell, J., Ward, C., Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L. G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., … Weiner, M. W. (2008). The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: JMRI*, *27*(4), 685–691.

Jazin, E., & Cahill, L. (2010). Sex differences in molecular neuroscience: from fruit flies to humans. *Nature Reviews. Neuroscience*, *11*(1), 9–17.

Jung, R. E., & Haier, R. J. (2007). The Parieto-Frontal Integration Theory (P-FIT) of intelligence: converging neuroimaging evidence. *The Behavioral and Brain Sciences*, *30*(2), 135–154; discussion 154–187.

Karrer, T. M., Bassett, D. S., Derntl, B., Gruber, O., Aleman, A., Jardri, R., Laird, A. R., Fox, P. T., Eickhoff, S. B., Grisel, O., Varoquaux, G., Thirion, B., & Bzdok, D. (2019). Brain-based ranking of cognitive domains to predict schizophrenia. *Human Brain Mapping*, *40*(15), 4487–4507.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., & Zisserman, A. (2017). The Kinetics Human Action Video Dataset. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1705.06950

Khan, N. M., Abraham, N., & Hon, M. (2019). Transfer Learning With Intelligent Training Data Selection for Prediction of Alzheimer's Disease. In *IEEE Access* (Vol. 7, pp. 72726–72735). https://doi.org/10.1109/access.2019.2920448

Kiesow, H., Dunbar, R. I. M., Kable, J. W., Kalenscher, T., Vogeley, K., Schilbach, L., Marquand, A. F., Wiecki, T. V., & Bzdok, D. (2020). 10,000 social brains: Sex differentiation in human brain anatomy. *Science Advances*, *6*(12), eaaz1170.

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1412.6980

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(13), 3521–3526.

Klein, S. L., Schiebinger, L., Stefanick, M. L., Cahill, L., Danska, J., de Vries, G. J., Kibbe, M. R., McCarthy, M. M., Mogil, J. S., Woodruff, T. K., & Zucker, I. (2015). Opinion: Sex inclusion in basic research drives discovery. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(17), 5257–5258.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Kunze, J., Kirsch, L., Kurenkov, I., Krug, A., Johannsmeier, J., & Stober, S. (2017). Transfer Learning for Speech Recognition on a Budget. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1706.00290

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A., & Denniston, A. K. (2019). A comparison of deep learning

performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet. Digital Health*, *1*(6), e271–e297.

Mateos-Pérez, J. M., Dadar, M., Lacalle-Aurioles, M., Iturria-Medina, Y., Zeighami, Y., & Evans, A. C. (2018). Structural neuroimaging as clinical predictor: A review of machine learning applications. *NeuroImage. Clinical*, *20*, 506–522.

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L. R., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., … Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, *19*(11), 1523–1536.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Peters, S. A. E., Huxley, R. R., & Woodward, M. (2014). Do smoking habits differ between women and men in contemporary Western populations? Evidence from half a million people in the UK Biobank study. *BMJ Open*, *4*(12), e005663.

Qin, C.-X., Qu, D., & Zhang, L.-H. (2018). Towards end-to-end speech recognition with transfer learning. *EURASIP Journal on Audio, Speech, and Music Processing*, *2018*(1), 1–9.

Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding Transfer Learning for Medical Imaging. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1902.07208

Rajpurkar, P., Park, A., Irvin, J., Chute, C., Bereket, M., Mastrodicasa, D., Langlotz, C. P., Lungren, M. P., Ng, A. Y., & Patel, B. N. (2020). AppendiXNet: Deep Learning for Diagnosis of Appendicitis from A Small Dataset of CT Exams Using Video Pretraining. *Scientific Reports*, *10*(1), 3958.

Ravishankar, H., Sudhakar, P., Venkataramani, R., Thiruvenkadam, S., Annangi, P., Babu, N., & Vaidya, V. (2017). Understanding the Mechanisms of Deep Transfer Learning for Medical Images. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1704.06040

Ritchie, S. J., Cox, S. R., Shen, X., Lombardo, M. V., Reus, L. M., Alloza, C., Harris, M. A.,

Alderson, H. L., Hunter, S., Neilson, E., Liewald, D. C. M., Auyeung, B., Whalley, H. C., Lawrie, S. M., Gale, C. R., Bastin, M. E., McIntosh, A. M., & Deary, I. J. (2018). Sex Differences in the Adult Human Brain: Evidence from 5216 UK Biobank Participants. *Cerebral Cortex* , *28*(8), 2959–2975.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Schulz, M.-A., Yeo, B. T. T., Vogelstein, J. T., Mourao-Miranada, J., Kather, J. N., Kording, K., Richards, B., & Bzdok, D. (2020). Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nature Communications*, *11*(1), 4238.

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1409.1556

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (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*(3), e1001779.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going Deeper with Convolutions. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1409.4842

Tannenbaum, C., Ellis, R. P., Eyssel, F., Zou, J., & Schiebinger, L. (2019). Sex and gender analysis improves science and engineering. *Nature*, *575*(7781), 137–146.

Taylor, B. K., Embury, C. M., Heinrichs-Graham, E., Frenzel, M. R., Eastman, J. A., Wiesman, A. I., Wang, Y.-P., Calhoun, V. D., Stephen, J. M., & Wilson, T. W. (2020). Neural oscillatory dynamics serving abstract reasoning reveal robust sex differences in typically-developing

children and adolescents. *Developmental Cognitive Neuroscience*, *42*, 100770.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A Closer Look at Spatiotemporal Convolutions for Action Recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6450–6459.

Wang, D., & Zheng, T. F. (2015). Transfer Learning for Speech and Language Processing. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1511.06066

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, *3*(1), 1–40.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 3320–3328.

Zawadzka-Gosk, E., Wołk, K., & Czarnowski, W. (2019). Deep Learning in State-of-the-Art Image Classification Exceeding 99% Accuracy. *New Knowledge in Information Systems and Technologies*, 946–957.