On the Utility of Large-Scale Pretrained Convolutional Neural Networks in Clinical Neuroimaging

Esten H. Leonardsen[1,2], Thomas Wolfers[3,4,1], Yunpeng Wang[1], and Lars T. Westlye[1,2,5]

[1]Department of Psychology, University of Oslo, Oslo, Norway
[2]Centre for Precision Psychiatry, Oslo University Hospital & Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[3]Department of Psychiatry and Psychotherapy, University of Tübingen, Tübingen, Germany
[4]German Center for Mental Health (DZPG), Jena, Germany
[5]KG Jebsen Center for Neurodevelopmental Disorders, University of Oslo, Oslo, Norway

December 15, 2025

1 Introduction

Over the last decade, deep learning has increasingly been applied to predict behavioural traits based on neuroimaging data. However, neuroimaging datasets containing well-characterised clinical phenotypes are often small, containing tens, hundreds, or, at best, thousands of participants, limiting the suitability of advanced modelling techniques for clinical prediction tasks. In other domains where deep learning has been successful, sample size issues are commonly alleviated using transfer learning. This technique builds on the intuition that a model that has successfully learned to solve a predictive task using a given data modality will necessarily encode knowledge that can be useful also when tackling related modelling problems.[1] In practice, this is commonly realised by pretraining a deep neural network using a large-scale dataset and a generic predictive target, and then later finetuning the model towards downstream tasks where data is scarcer. In general, the application of transfer learning has been shown to broadly increase the applicability of deep learning in domains where the amount of data should theoretically impede their use.[2] Still, the adoption of transfer learning in clinical neuroimaging studies is limited. We

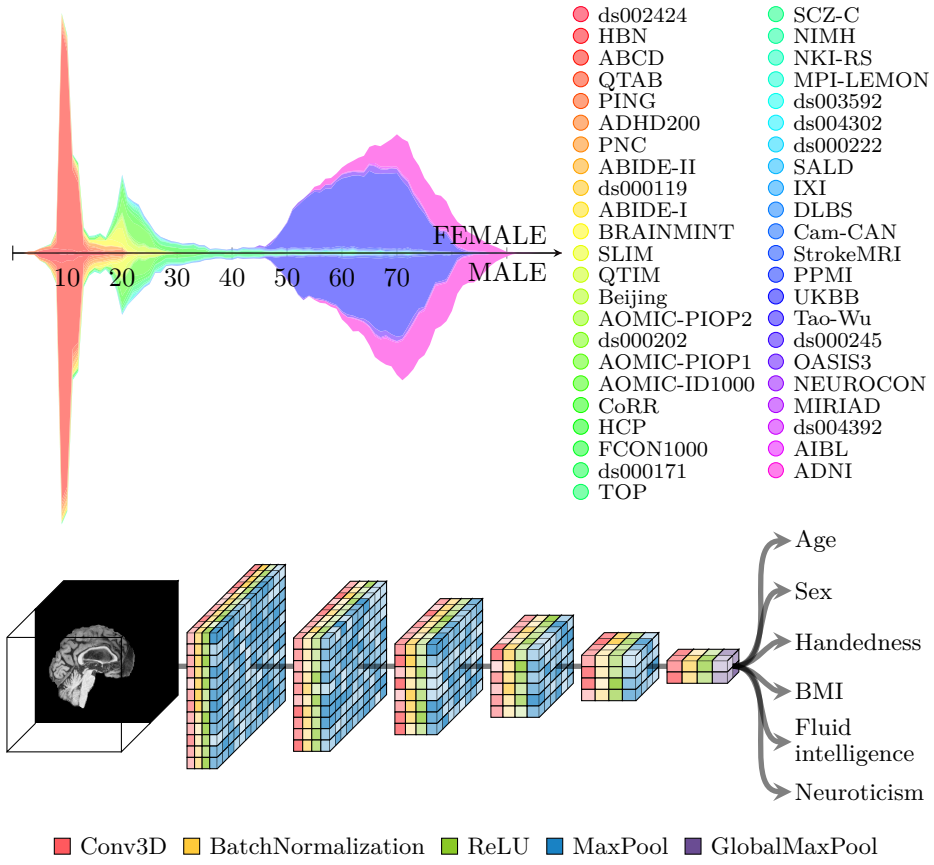
hypothesize that this is in part due to the limited availability of pretrained neuroimaging models trained with sufficient amounts of data. Furthermore, it has been revealed that a common pretraining approach underlying several of the models that do exist does not yield the expected performance gains,[3] limiting their downstream utility. In this work we address these issues by compiling a large dataset (n=114,257) of structural T1-weighted magnetic resonance images (MRI) to pretrain a convolutional neural network (CNN) and investigate its usefulness for downstream modelling tasks.

2 Methods

The full dataset used in the study was compiled from 45 different cohorts (114,257 images from 80,931 participants, age range=3-97 years, 49% females). The age distribution of the dataset and an overview of the cohorts can be seen in Figure 1. This dataset was split into three parts: a training set (68,570 images from 59,923 participants) used for pretraining, a validation set (10,000 images from 10,000 participants) used for epoch selection during pretraining, and a test set (34,250 images from 13,442 participants) used for testing the efficacy of the pretrained model in downstream tasks. The first two sets were drawn from 25 cohorts containing only healthy individuals, whereas the last was drawn from 20 disjoint cohorts not seen by the model during pretraining (with the exception of data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), where some individuals were used for pretraining to ensure the age-range spanned that of the test set, and the remaining were used for downstream tasks). Before modelling, each MRI was minimally preprocessed using FastSurfer,[4] to ensure a uniform orientation and remove non-brain tissue.

The model was a variant of the Simple Fully Convolutional Network (SFCN)-architecture,[5] a small 3-dimensional CNN consisting of five repeated convolutional blocks and a downsampling block. On top of this backbone, we constructed a multi-task prediction head with six output neurons to simultaneously predict age, sex, handedness, body mass index (BMI), fluid intelligence and neuroticism. During pretraining, a weighted loss was calculated across all these predictive targets, while masking out missing entries, resulting in a multi-task model.

To test the usefulness of transferring the pretrained model to downstream tasks, we constructed two clinical classification problems, each accompanied by a tailored pool of data. The first task involved differentiating patients with Alzheimer's disease (AD) and healthy controls (HC), relying on 10,660 images from 1,678 participants from the ADNI dataset. The second task consisted of classifying patients with schizophrenia (SCZ) and bipolar disorder (BD), using 1,037 images from 861 participants from the Thematically Organized Psychosis (TOP) dataset. For each task and $n \in \{100, 200, 300, 400, 500\}$, we first sampled n images with replacement from the pool to use for training binary classifiers. Next we sampled $\frac{n}{5}$ images to use as a validation set and a test set of $\frac{n}{5}$ images to measure predictive performance using the AUC in a held-out dataset,



enforcing that images from a single participant did not appear in multiple sets. For each sampled dataset, we fit three types of models: logistic regression models using regional volumes from FastSurfer (referred to as linear models below), a binary SFCN-variant with randomly initialized weights (referred to as the baseline model), and a binary SFCN initialized with the pretrained weights from the multi-task model (referred to as the transfer learning model). For each model type we fit 9 models with different hyperparameter settings for each sampled dataset and selected the best model using the validation set. This process was repeated 100 times per n per task, resulting in 100 out-of-sample AUCs per model type. Differences in pairwise model performance was assessed by calculating whether the AUCs of one model type was consistently greater than the AUCs of another using a paired permutation test with a predefined significance threshold of 0.05.

3 Results

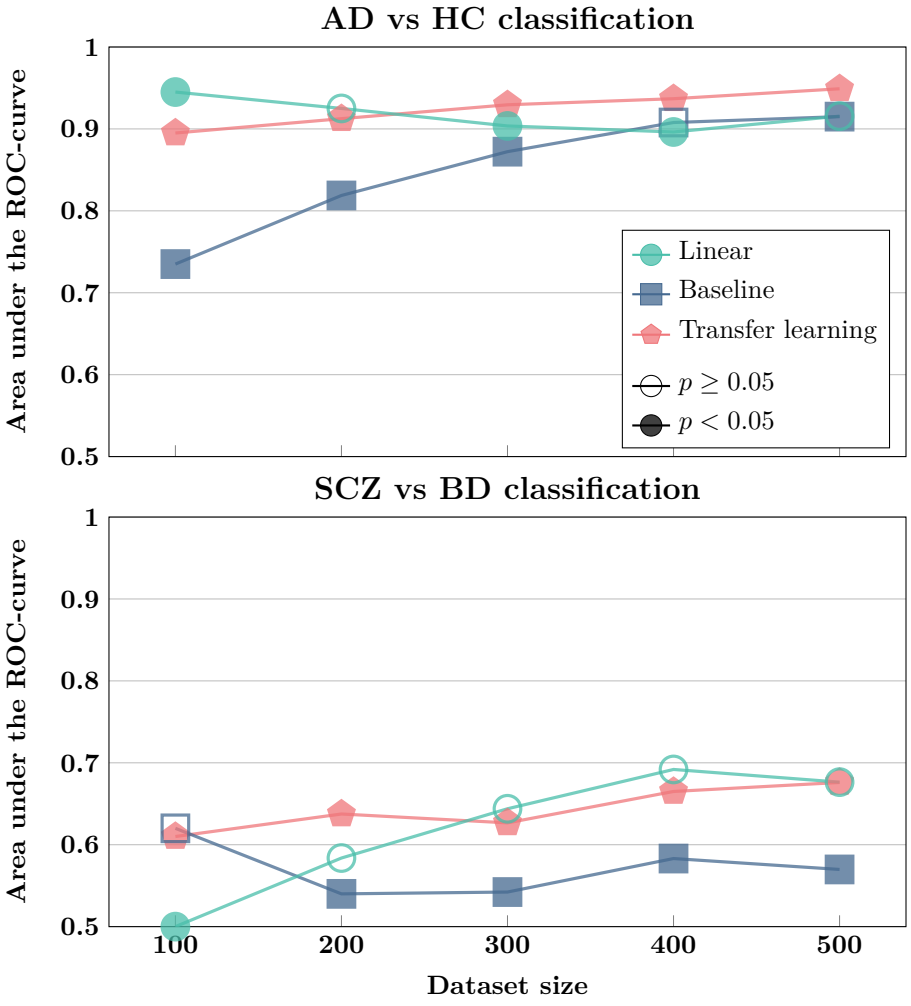
In the validation set, the pretrained multi-task model resulted in accurate predictions for age (MAE=2.23) and sex (AUC=0.99), and more modest performance for BMI (MAE=2.38, compared to MAE=5.6 from a naïve dummy model), handedness (AUC=0.61), fluid intelligence (R=0.27) and neuroticism (R=0.17). The model also achieved high performance for predicting age (MAE=3.35) and sex (AUC=0.98) in the held-out test set, consisting of images from scanners and cohorts not seen during training.

For the AD vs HC classification task, the overall worst performance was achieved by the baseline CNN model, with median AUCs that monotonically increased from 0.72 for the smallest sample size (n=100) to 0.92 for the largest sample size (n=500). For n=100, the best performance was achieved by the linear model, reaching a median AUC of 0.95. This significantly outperformed ($p<0.05$ across the permutation test) the median AUC of 0.86 achieved by the transfer learning model. For n=200, the linear model seemingly also outperformed the transfer learning model, with a median AUC of 0.93 compared to 0.89, although this difference did not reach statistical significance. For the three largest dataset sizes ($n \in \{300, 400, 500\}$) the transfer learning model significantly outperformed both the baseline and the linear model, reaching a maximum median AUC of 0.94 (n=500).

For the SCZ vs BD classification task, the worst performance for $n \in \{200, 300, 400, 500\}$ was once again achieved by the baseline CNN, with median AUCs spanning 0.54 to 0.58. However, somewhat surprisingly, this model outperformed both the other models for n=100 with a median AUC of 0.62, although the improvement over the transfer learning model did not reach statistical significance. The transfer learning model achieved median AUCs of 0.61 to 0.68, with a positive, although non-monotonic, trend in performance as samples sizes grew. For n=200, this model type significantly outperformed both other models in the permutation test. The linear model saw a steeper improvement with larger sample sizes, spanning median AUCs from 0.5 (n=100) to 0.69 (n=400), although neither this was monotonic (median AUC=0.68 for n=500). For the three largest sample sizes, the differences in performance between the transfer learning model and the linear model were statistically indistinguishable.

4 Conclusion

Our results demonstrate that pretraining a CNN in a large-scale neuroimaging dataset prior to training towards new tasks in smaller clinical datasets yields consistent performance improvements. Furthermore, they suggest that a multi-task pretraining approach might overcome some of the challenges that have been reported for more common single-task pretraining methods. Nonetheless, they also show that relatively simple linear models based on volumetric inputs are competitive with advanced CNNs when the number of training samples are in


AD vs HC classification
Area under the ROC-curve
1
0.9
0.8
0.7
0.6
0.5
Linear
Baseline
Transfer learning
$p \geq 0.05$
$p < 0.05$
SCZ vs BD classification
Area under the ROC-curve
1
0.9
0.8
0.7
0.6
0.5
100
200
300
400
500
Dataset size

the hundreds.

References

1. Razavian AS, Azizpour H, Sullivan J, Carlsson S. *CNN Features off-the-shelf: an Astounding Baseline for Recognition*. 2014. arXiv: 1403.6382 [cs.CV].
2. Morid MA, Borjali A, Del fiol G. A scoping review of transfer learning research on medical image analysis using ImageNet. *Computers in Biology and Medicine*. 2021; 128():104115.
3. Tan TWK, Nguyen KN, Zhang C, et al. Mind the Gap: Does Brain Age Improve Alzheimer's Disease Prediction? *Human Brain Mapping*. 2025; 46(12):e70276.
4. Henschel L, Conjeti S, Estrada S, Diers K, Fischl B, Reuter M. FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*. 2020; 219():117012.
5. Peng H, Gong W, Beckmann CF, Vedaldi A, Smith SM. Accurate brain age prediction with lightweight deep neural networks. *Medical image analysis*. 2021; 68():101871.