Sex classification from functional brain connectivity: Generalization to multiple datasets

Generalizability of sex classifiers

Lisa Wiersch[1,2], Patrick Friedrich[1,2], Sami Hamdan[1,2], Vera Komeyer[1,2], Felix Hoffstaedter[1,2], Kaustubh R. Patil[1,2], Simon B. Eickhoff[1,2] and Susanne Weis[1,2]

[1]Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[2]Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany

Corresponding author: s.weis@fz-juelich.de

Abstract

Machine learning (ML) approaches are increasingly being applied to neuroimaging data. Studies in neuroscience typically have to rely on a limited set of training data which may impair the generalizability of ML models. However, it is still unclear which kind of training sample is best suited to optimize generalization performance. In the present study, we systematically investigated the generalization performance of sex classification models trained on the parcelwise connectivity profile of either single samples or a compound sample containing data from four different datasets. Generalization performance was quantified in terms of mean across-sample classification accuracy and spatial consistency of accurately classifying parcels. Our results indicate that generalization performance of pwCs trained on single dataset samples is dependent on the specific test samples. Certain datasets seem to "match" in the sense that classifiers trained on a sample from one dataset achieved a high accuracy when tested on the respected other one and vice versa. The pwC trained on the compound sample demonstrated overall highest generalization performance for all test samples, including one derived from a dataset not included in building the training samples. Thus, our results indicate that a big and heterogenous training sample comprising data of multiple datasets is best suited to achieve generalizable results.

Keywords: machine learning, sex classification, generalizability, resting-state functional connectivity, neuroimaging, big data

Introduction

Machine Learning (ML) is a powerful tool to relate neuroimaging data to behavior and phenotypes (Genon et al., 2022; Varoquaux & Thirion 2014) and is therefore increasingly being employed in neuroscience applications (Jollans et al., 2019; Buch et al., 2018; Varoquaux et al., 2018; Kohoutova et al. 2020). Successful applications of ML approaches include the decoding of mental states (Haynes & Rees 2006), classification of mental disorders (Zhang et al. 2021; Chen et al., 2020), as well as the prediction of demographic and behavioral phenotypes (Smith et al. 2015; Nostro et al., 2018; Pläschke et al., 2020; Varikuti et al., 2018; More et al., 2023).

ML models learn the feature-target relationship given a training sample. Subsequently, the model is applied to make predictions on previously unseen data (Dhamala et al., 2023) and successful generalization to independent data samples is the central goal in ML (Domingos, 2012; Varoquaux, 2018; Chung, 2018). For example, a recent study (Weis et al., 2020) demonstrated successful generalization of sex prediction models based on regionally specific functional brain connectivity patterns, which were trained on the data of the Human Connectome Project (HCP, Van Essen et al., 2012, Van Essen, 2013). For this spatially specific approach, independent classifiers were trained on the functional brain connectivity patterns of parcels covering the whole brain. In this case, assessing generalization performance should not only consider the averaged across-sample accuracy. Rather, if the classifiers generalize well, the same parcels should achieve high classification accuracies during cross-validation (CV) and across-sample testing.

Further sex classification studies (Menon & Krishnamurthy, 2019; Zhang et al., 2018; Smith et al., 2013), as well as other applications of ML models employed the HCP dataset to predict phenotypes such as task activation (Cohen et al., 2020), and individual behavioral and demographic scores (Smith et al., 2015; Cui & Gong, 2018) like age (Sanford et al., 2022). The HCP dataset is characterized by high-quality multi-modal imaging data acquired from a large group of healthy young adults. However, both the high quality of the brain imaging data as well as the narrow age range is not typical of other datasets, especially when dealing with clinical data (Arslan, 2018, Jansma et al., 2020; Rutten et al., 2010). This raises the question whether results based on the HCP data can be generalized to other datasets with different characteristics. Weis et al., (2020) demonstrated that sex classifiers trained on the HCP data generalized well to an independent subset of the HCP dataset as well as to the 1000Brains dataset (Caspers et al. 2014). Additional evidence from the application of such classifiers to data from datasets with diverse characteristics would provide even stronger evidence of model generalization.

Especially in neuroimaging, differences between datasets may result from several different sources. On the one hand, participants may differ with respect to demographic characteristics, such as age, education, or economic status. On the other hand, data samples likely differ with regard to the MRI acquisition parameters and data processing. Considering these differences, it is so far unresolved what kind of training sample leads to good generalization performance across multiple test samples.

Various characteristics of the training data can influence the generalization performance of ML models (Dhamala et al. 2023). For instance, larger sample size is beneficial for generalization performance (Cui & Gong, 2018, Domingos, 2012). Ensuring that the training data is representative of the target sample is another crucial factor for achieving good generalization performance (Ishida, 2019, Yang et al. 2020). Furthermore, data from different acquisition sites are likely heterogeneous with respect to demographic characteristics, data acquisition, and processing parameters. Therefore, a ML model trained on such data is less likely to overfit. Thus, aggregating data from multiple sites should be beneficial for improving generalization performance. Indeed, this has been partially shown by studies concerning clinical applications of ML approaches (Nielsen et al, 2020; Willemink et al. 2020; Chang et al. 2018). These results suggest that training ML models on diverse datasets covering a wide range of characteristics may improve the overall generalization performance.

In the present study, our aim was to evaluate the generalization performance of sex classifiers trained on samples created from four different datasets with varying demographic characteristics. In addition, sex classifiers were trained on a compound sample combining data from all datasets to obtain a training sample with heterogeneous sample characteristics. Following the parcelwise approach by Weis et al. (2020), we trained independent sex classifiers on the resting state (RS) connectivity patterns of 436 parcels covering the whole brain. For each parcel, a sex classification model was built based on the individual connectivity profile, resulting in one classification accuracy value per parcel. This was done for each of the five training samples, resulting in five sets of parcelwise classifiers (pwCs). These pwCs were applied to test samples from the four original datasets and one dataset which was not part of the training samples. Then, accuracy maps, representing the spatial distribution of classification accuracies for each parcel were generated for CV (within-sample accuracy) and for application of the pwCs to the different test samples (across-sample accuracy). The comparison of these accuracy maps enabled us to evaluate generalization performance of classifiers by (i) examining the mean accuracy of all parcelwise classifiers across the 10% best classifying parcels and (ii) comparing the spatial location of highly classifying parcels between CV and across-sample test. Good generalization performance with regard to spatial consistency is characterized by identical parcels performing well in CV and across-sample testing. We hypothesized that the compound sample achieves best generalization performance as suggested by previous literature (Nielsen et al, 2020; Willemink et al. 2020; Chang et al. 2018).

Materials and Methods

Data

We employed resting state functional magnetic resonance imaging (fMRI) data of subsets of four large datasets to train and test sex classification models. For all datasets, we only included healthy subjects aged 20 years or older. Within each training sample, we matched females and males for age and included a similar number of women and men. The first sample, taken from the HCP dataset (900 subjects data release; Van Essen et al., 2012; Van Essen 2013), comprised 878 subjects with a mean age of 28.49 years (range: 22-37 years). The second sample, taken from the Brain Genomics Superstruct Project (GSP; Holmes et al., 2015) comprised 854 subjects with a mean age of 22.92 years (range: 21 – 35 years). The third sample was a subset from the Rockland Sample of the Enhanced Nathan Klein Institute (eNKI; Nooner et al., 2012), comprising 190 subjects with a mean age of 46.02 years (range: 20-83 years). The fourth sample, taken from the 1000Brains dataset (Caspers et al., 2014), comprised 1000 subjects with a mean age of 61.18 years (range: 21-85 years). This sample was included to examine generalization performance to an older sample. A fifth sample ("compound") was constructed by combining 75% of the HCP, GSP, eNKI and 1000Brains samples. The compound sample comprised an age range of 20-85 years (*M* = 40.10, *SD* = 19.96 years). RS fMRI data from an additional dataset was included to evaluate classifiers on an additional independent sample. This sample comprised 370 subjects (214 females) with a mean age of 22.50 years (range 20-26 years) from the AOMIC dataset (Snoek et al., 2021). It was not additionally balanced for sex to maintain the maximum number of participants for evaluation. Data usage of the included datasets was approved by the Ethics Committee of the Medical Faculty of the Heinrich-Heine University Düsseldorf (4039, 5193, 2018-317-RetroDEuA). All data was collected in research projects approved by a local Review Board, for which all participants provided written informed consent. All experiments were performed in accordance with relevant guidelines and regulations.

Data acquisition and preprocessing

HCP

The RS fMRI data of the HCP dataset were acquired on a Siemens Skyra 3T MR scanner with multiband echo-planar imaging with a duration of 873 seconds and the following parameters: 72 slices; voxel size, 2 x 2 x 2 mm^3; field of view (FOV), 208 x 180 mm^2; matrix, 104 x 90; TR, 720 ms; TE, 33 ms; flip angle, 52 degrees (https://www.humanconnectome.org/storage/app/media/documentation/s1200/HCP_S1200_Release_Reference_Manual.pdf) Participants were instructed to lie in the scanner with eyes open, with a "relaxed" fixation on a white cross on a dark background and think of nothing in particular, and to not fall asleep (Smith et al., 2013). RS data were corrected for spatial distortions, head motion, B$_0$ distortions and were registered to the T1-weighted structural image (Smith et al. 2013). Concatenating these transformations with the structural-to-MNI nonlinear warp field resulted in a single warp per time point, which was applied to the timeseries to achieve a single resampling in the 2mm MNI space. Afterwards, global intensity normalization was applied and voxels that were not part of the brain were masked out. Locally noisy voxels as measured by the coefficient of variation were excluded and all the data were regularized with 2mm Full width half maximum (FWHM) surface smoothing (Smith et al. 2013; Glasser et al. 2013). The temporal preprocessing included corrections and removal of physiological and movement artifacts by an independent component analysis (ICA) of the FMRIB´s X-noisifier (FIX, Salimi-Khorshidi et al., 2014). This method decomposes data into independent components and identifies noise components based on a variety of spatial and temporal features through pattern classification.

GSP

RS data were acquired on a 3T Tim Trio Scanner with a duration of 372 seconds and the following parameters: 47 slices; voxel size, 3 x 3 x 3 mm^3; FOV read, 216 mm; TR, 3 s; TE, 30 ms; flip angle, 85 degrees. During data acquisition, participants were instructed to lay still, stay awake, and keep eyes open while blinking normally (https://static1.squarespace.com/static/5b58b6da7106992fb15f7d50/t/5b68650d8a922db3bb807a90/1533568270847/GSP_README_140630.pdf, Holmes et al. 2015).

eNKI

Participants in the eNKI dataset were underwent RS scanning for 650 seconds in a Siemens Magnetom Trio Tim sygno MR scanner with the following parameters: 38 slices; voxel size, 3 x 3 x 3 mm^3, FOV, 256 x 200mm^2; TR, 2500 ms; TE, 30 ms; flip angle, 80 degree. Participants were instructed to keep their eyes closed, relax their minds and not to move (Betzel et al. 2014).

1000Brains

Subjects were scanned for 660 seconds on a Siemens TRIO 3T MRI scanner with the following parameters: 36 slices; voxel size, 3.1 x 3.1 x 3.1 mm^3; FOV, 200 x 200 mm^2; matrix, 64 x 64, TR = 2.2 s; TE = 30 ms; flip angle, 90 degrees. During RS data acquisition, participants were instructed to keep their eyes closed and let the mind wander without thinking of anything in particular (Caspers et al. 2014).

RS data of the GSP, eNKI and 1000Brains samples were preprocessed in the same way. The preprocessing pipeline comprised removal of noise and motion artifacts by using FIX (Salimi-Khorshidi et al., 2014). The denoised data were further preprocessed with SPM12 (SPM12 v6685, Wellcome Centre for Human Neuroimaging, 2018; https://www.fil.ion.ucl.ac.uk/spm/software/spm12/) using Matlab R2014a (Mathworks, Natick, MA). For each subject, the first four echo-planar-imaging (EPI) volumes were discarded and the remaining ones were corrected for head movement by an affine registration with two steps: First, the images were aligned to the first image. Second, the images were aligned to the mean of all volumes. The resulting mean EPI image was spatially normalized using the MNI152 template (Holmes et al., 1998) using the “unified segmentation” approach in order to take into account inter-individual differences in brain morphology (Ashburner & Friston, 2005). After deformation application and smoothing, images were resampled to partial volume label image to 2x2x2mm^3 and resampled using the modulated GM segment image to 2x2x2mm^3.

AOMIC

The AOMIC dataset includes two subsamples, PIOP1 and PIOP2, comprising data of healthy university students scanned on a Philips 3T scanner. Participants were instructed to keep their gaze fixated on a fixation cross on the screen and let their thoughts run freely (Snoek et al., 2021). Both samples were acquired with a voxel size of 3 x 3 x 3 mm^3 and a matrix size of 80 x 80. While PIOP1 was acquired for 360 seconds with multi-slice acceleration, 480 volumes and a 0.75 TR, PIOP2 was acquired for 480 seconds without multi-slice acceleration, 240 volumes and a 2s TR (further details in https://www.nature.com/articles/s41597-021-00870-6/tables/10). Data were preprocessed using Fmriprep version 1.4.1 (Esteban et al., 2019; Esteban et al., 2020), a Nipype based tool for reproducible preprocessing in neuroimaging data (Gorgolewski et al., 2011). Data were motion corrected using mcflirt (FSLv5.0.9, (Jenkinson et al. 2002)) followed by distortion correction by co-registering the functional image to the respective T1 weighted image with inverted intensity (Huntenburg, 2014; Wang et al., 2017) with 6 degrees of freedom, using bbregister (FreeSurfer v6.0.1). In a following step, motion correction transformations, field distortion correction warp, BOLD-to-T1- weighted transformation and the warp from T1-weighted to MNI were concatenated and applied using antsApplyTransforms (ANTs v2.1.0.) using Lanczos interpolation (Snoek et al., 2021).

Connectome extraction

Following the parcelwise approach by Weis et al. (2020), individual RS connectomes were extracted based on 400 cortical parcels of the Schaefer Atlas (Schaefer et al. 2018), and 36 subcortical parcels of the Brainnetome Atlas (Fan et al., 2016). Each parcel's time series was cleaned by excluding variance that could be explained by mean white matter and cerebrospinal fluid signal (Satterthwaite et al., 2013). Data was not further cleaned for motion related variance as this variance was already removed during FIX preprocessing. For each of the 436 parcels, the activation time series was computed as the mean of all voxel time courses within that parcel. Then, for each parcel, pairwise Pearson correlations were computed between the parcel’s time series and those of all other 435 remaining parcels, representing the individual RS functional connectivity (RSFC) profile of the parcel.

Parcelwise sex classification

Sex classification models were trained based on the individual multivariate RSFC profile of each parcel, resulting in a set of 436 pwC (Weis et al., 2020). All models were built using support vector machine (SVM) classifiers. SVM is a supervised ML method that separates the data into distinct classes – males and females in case of sex classification – with the widest possible gap between these classes (Vapnik 1998; Boser et al., 1992; Rafi & Shaikh., 2013; Zhang et al., 2021). SVM models were built in Julearn (https://juaml.github.io/julearn/main/index.html) including a Hyperparameter search nested within a 10–fold CV with 5 repetitions. The parameter search included choice of kernel (linear vs. radial basis function (rbf) kernel) as well as the C- and gamma-parameter. Confounding effects of age were regressed out by removing age-related variance before training the classifiers. The best performing combination of hyperparameters was used for the final model for each individual parcel. Within-sample classification accuracy for each individual parcel was determined by averaging accuracies over CV folds and repetitions.

For across-sample classification, single dataset pwCs were tested on the respective other three samples (sample characteristics displayed in table 1, Figure 1), while the compound sample pwC were tested on the remaining 25% of the HCP (n = 220, age range: 22-36), GSP (n = 214, age range: 21-31), eNKI (n = 48, age range: 20-75) and 1000Brains (n = 250, age range: 22-80) sample (Table 1). Here, for computing time reasons, we restricted the choice of the SVM kernel to rbf (see Weis et al., 2020). Finally, generalization performance of all five pwCs was assessed on the AOMIC sample. All reported accuracies are balanced accuracies.

Statistical analyses

Across-sample classification accuracy

To statistically compare the classification accuracies achieved by each pwC on the different test samples, we employed independent t-tests between the different across-sample accuracies over all 436 parcels. To compare the performance of the different pwCs on each test sample, independent t-tests across all parcels were used. Significance levels were Bonferroni-corrected according to the number of dependent tests (10 dependent tests for comparing the across-sample accuracy of pwC compound on the five fest samples as well as for comparing the across-sample accuracies of the five pwCs on the AOMIC sample; 6 dependent tests for all other comparisons).

Consistency of highly classifying brain regions

Previous studies have demonstrated that sex classification accuracies for models trained on parcelwise RSFC patterns do not achieve uniformly high performance across the whole brain (Weis et al. 2020; Zhang et al. 2018). Thus, we assessed generalization performance of the different pwCs by examining the consistency of highly classifying brain regions during CV and across-sample testing. Consistency was assessed by computing Dice coefficients (DSC) to evaluate the similarity in spatial distribution of parcels achieving certain accuracies in both CV and across-sample testing. This consistency was evaluated for different accuracy thresholds above chance (0.5 - 0.7 at 0.02 steps). For each threshold, Dice coefficients were computed as the number of common parcels achieving within- and across-sample accuracies above or equal to that threshold (p_com) multiplied by 2 and divided by the total number of parcels achieving a within (p_tr) or across-sample (p_te) accuracy above or equal to that accuracy level in CV (Dice, 1945; Sorensen, 1948).

$$DSC = \frac{2 * p\_com}{p\_tr + p\_te}$$

To facilitate comparison of the dice score distributions between the different pwCs and test samples, we summarized each contribution into one score by computing a weighted mean (wmDice) as the average of each dice coefficient weighted by the accuracy threshold for which the respective dice coefficient was calculated.

Table 1. Sample characteristics. Depiction of sample characteristics of each sample to train the respective pwC.

	pwC HCP	pwC GSP	pwC eNKI	pwC 1000Brains	pwC compound
training sample	HCP	GSP	eNKI	1000Brains	75% of the HCP, GSP, eNKI & 1000Brains sample
sample size	878	854	190	1000	2190
mean age (age range) years	28.49 (22-37)	22.92 (21-35)	46.02 (20-83)	61.18 (21-85)	40.10 (20-85)

Results

The generalization performance of pwCs trained on each of the single dataset samples (HCP, GSP, eNKI, & 1000Brains) and on the compound sample were compared with respect to mean across-sample accuracy averaged across the best 10% classifying parcels. Additionally, we evaluated the consistency of the spatial distribution of accurately classifying parcels between CV and across-sample testing.

Training and test classification accuracies

For the single samples pwCs, the mean within-sample performance across the top 10% classifying parcels was at a similar level for pwC GSP (66.8%), pwC eNKI (66.9%) and pwC 1000Brains (66.3%) and ranged up to 73.5% for pwC HCP. The mean across-sample accuracies averaged for the top 10% classifying parcels ranged between 58.4% (for pwC HCP tested on AOMIC and pwC eNKI tested on 1000Brains) and 65.8% (for pwC GSP tested on eNKI). Details for within- and across-sample performance are reported in Table S1 and Figure 1 and Figure S1. Parcelwise within- and across-sample accuracies are displayed as accuracy maps in figure 1a and the distribution of test accuracies is shown in figure 2 (red boxplots). Here, accuracy maps represent the spatial distribution of classification accuracies resulting from the 436 individual ML models trained on the respective multivariate RSFC profile of each parcel.

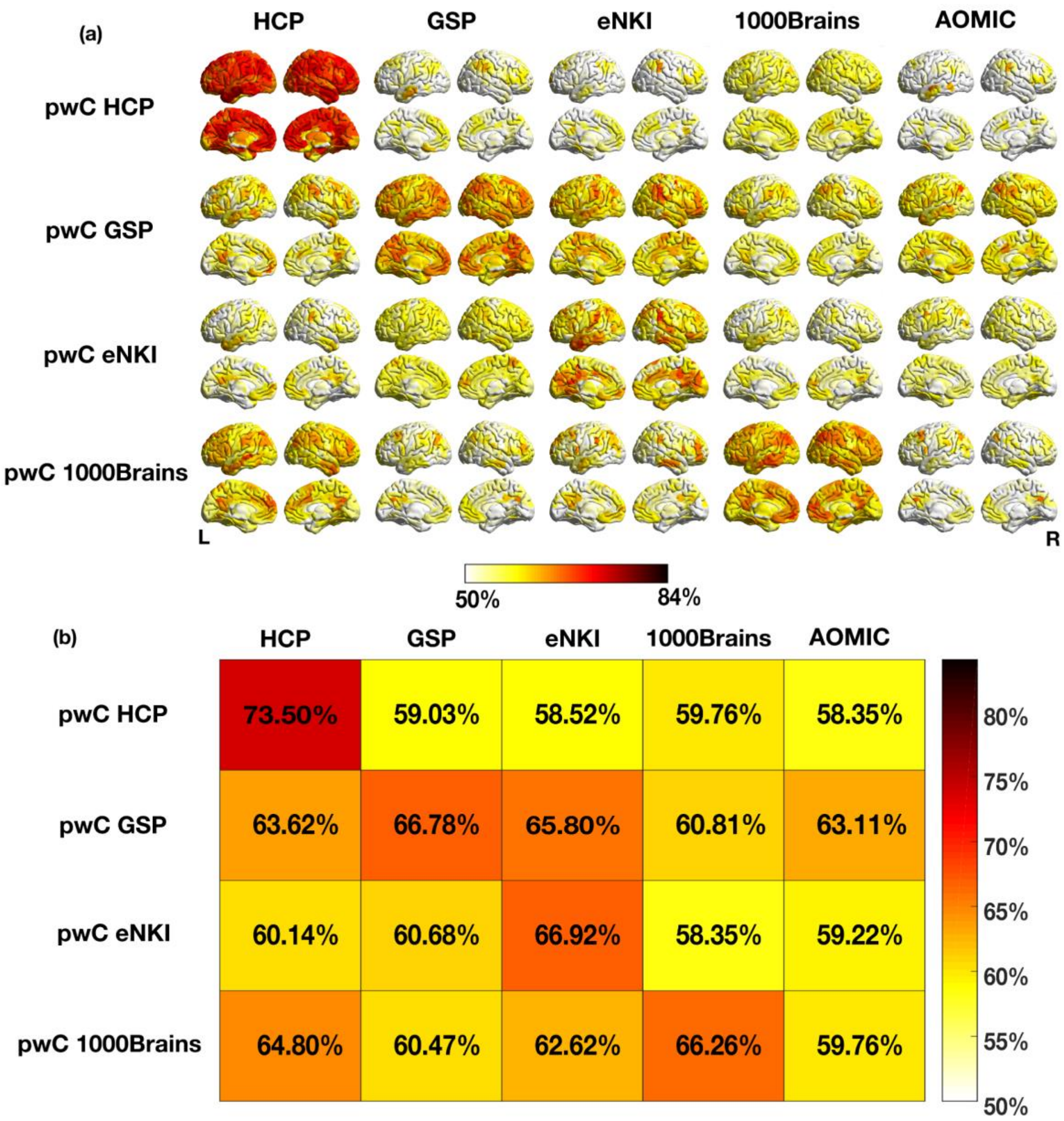

Figure 1. Accuracy maps and tile plots of mean accuracies of top 10% classifying parcels for pwCs trained on single samples.
(a) Spatial distribution of parcelwise sex classification accuracies across the brain. Within-sample accuracies are depicted on and across-sample accuracies off the diagonal. Only parcels with an accuracy of 0.5 or higher are displayed. (b) Mean accuracies averaged across the top 10% classifying parcels for each CV- and across-sample prediction.

Accuracy maps for the different combinations of training and test samples were compared using independent t-tests across the top 10% classifying parcels in each prediction (details in Table S2). First, we analyzed differences in classification accuracies between test samples for each pwC (horizontal comparisons, figure 1): For pwC HCP, testing on 1000Brains achieved the highest mean classification accuracy (59.8%). The averaged accuracy for this test sample was descriptively higher than for the GSP and significantly higher than for the eNKI and AOMIC test samples. PwC GSP achieved significantly higher accuracies for the eNKI test sample (65.8%) than for any other test sample, while pwC eNKI showed highest accuracies for the GSP test sample (60.7%). This across-sample prediction showed descriptively higher accuracies than pwC eNKI did for the HCP test sample and significantly higher accuracies than for the AOMIC and 1000Brains samples. For pwC 1000Brains, testing on the HCP showed significantly higher accuracies (64.8%) than testing on the eNKI, GSP and AOMIC sample. Details of all statistical comparisons are given in Table S2.

PwC compound achieved a mean within-sample accuracy of 67.9% within the top 10% classifying parcels. The mean across-sample accuracies averaged across the top 10% classifying parcels ranged between 65.5% (pwC compound tested on AOMIC) and 74.6% (pwC compound tested on eNKI, details in Table S1 and Figure 2, Figure S2).

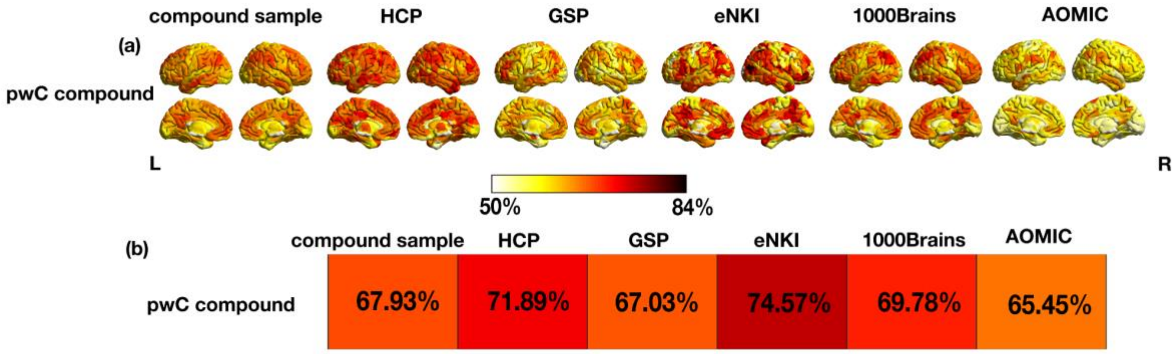

Figure 2. Accuracy maps and tile plots of mean accuracies of top 10% classifying parcels for pwC compound. (a) Spatial distribution of parcelwise sex classification accuracies across the brain. Only parcels with an accuracy of 0.5 or higher are displayed.
(b) Mean accuracies averaged across the top 10% classifying parcels for the CV- and across-sample prediction.

Contrasting the top 10% classifying parcels in the accuracy maps of pwC compound displayed peaks in accuracies for the eNKI test sample (74.6%) resulting in significantly higher accuracies than for the remaining test samples (Figure 2 and Table S2). We also contrasted how the five pwCs performed on each test sample by employing independent t-tests: pwC compound outperformed all pwCs trained on single dataset samples for the HCP, GSP, eNKI and AOMIC test sample with regards to the top 10%classifying parcels in each across-sample prediction. Details for all statistical comparisons are shown in Table S2.

Consistency of correctly classifying parcels

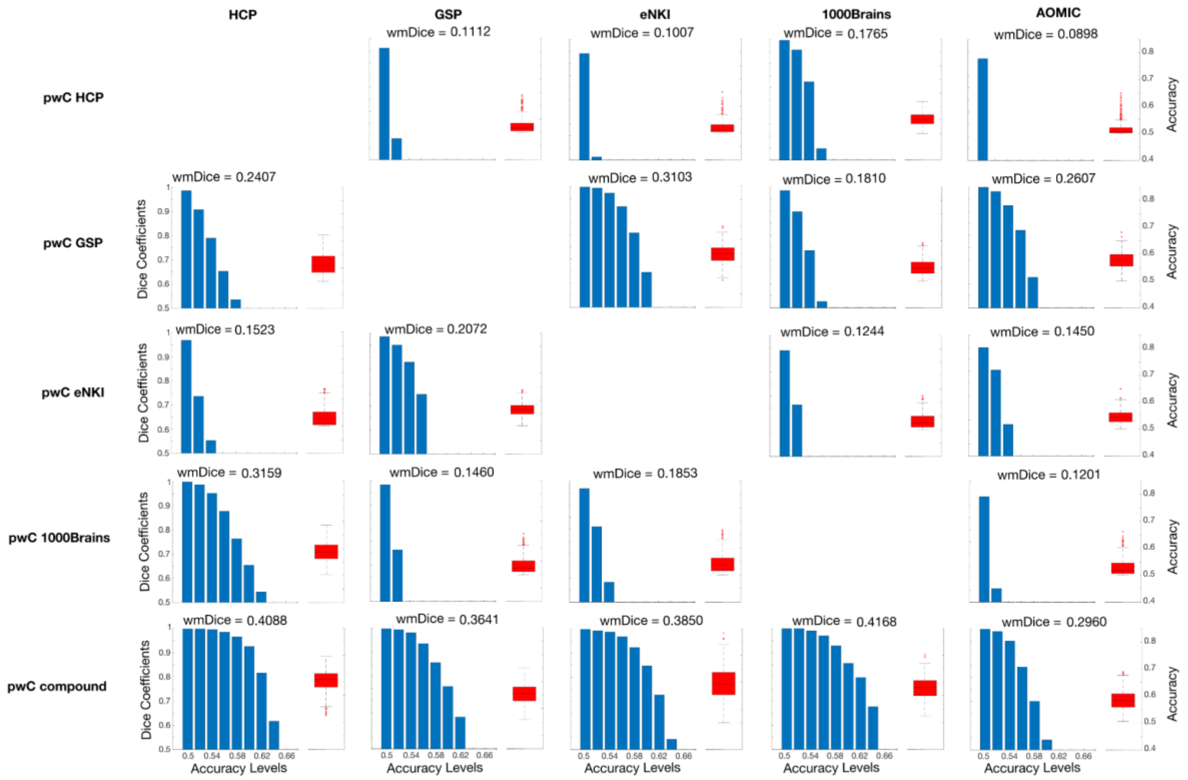

Figure 2. Spatial consistency of all pwCs. For each combination of training (rows) and test sample (columns), the right side of each subplot (red boxplot) depicts the distribution of accuracies across all parcels (right y-axis). The left side of each subplot (blue barplot) shows the dice coefficients (left y-axis), representing the overlap of accuracy maps between CV and test at different accuracy levels (x-axis). For each accuracy-threshold, the respective dice coefficient was calculated as the number of similar parcels classifying above a certain accuracy-threshold in both, respective CV and test prediction, in relation to the total number of parcels of both predictions classifying at this level. For each combination of pwC and test sample, the weighted mean of the dice coefficients (wmDice) across accuracy levels is displayed above the subplot to allow for a straightforward comparison between the distributions of dice coefficients.

To evaluate the spatial consistency of accurately classifying parcels, we calculated the dice coefficient between thresholded within- and across-sample accuracy maps at different levels of accuracy. Here, a high dice coefficient indicates a high overlap in highly classifying parcels between within and across-sample predictions at a given accuracy level. The results are depicted in the blue bar plots in figure 2. Regarding spatial consistency within a given pwC (horizontal comparison in Fig 2), pwC HCP overall demonstrated relatively low spatial consistency while it was highest for 1000Brains (wmDice = 0.1765, all other wmDice < 0.1112). Spatial consistency for pwC GSP was highest for the eNKI sample (wm = 0.3103) and lowest for 1000Brains (wmDice = 0.1810) with spatial consistency for HCP (wmDice = 0.2407) and AOMIC (wmDice = 0.2607) test samples ranging in between. PwC eNKI showed overall low spatial consistency for the HCP, 1000Brains and AOMIC sample (wmDice: 0.1244 - 0.1523) and highest for the GSP sample (wmDice = 0.2072). Spatial consistency of pwC 1000Brains was lower for the GSP, eNKI and AOMIC test sample (wmDice: 0.1201 - 0.1853) but considerably higher for the HCP test sample (wmDice = 0.3159) pwC compound demonstrated relatively similar spatial consistency for HCP, GSP, eNKI and 1000Brains (wmDice: 0.3641 - 0.4168) and lower spatial consistency with the AOMIC sample (wmDice: 0.2960). Concerning the comparisons within each test sample (vertical comparisons in Fig 2) pwC compound demonstrated higher spatial consistency than all single dataset sample pwCs for all test samples. Dice coefficients for the top 10% classifying parcels are reported in figure S3.

Discussion

In the present study, we examined the generalization performance of parcelwise sex classification models trained on different samples. Here, we operationalized generalization performance in terms of both mean classification accuracy of best classifying parcels during across-sample testing as well as spatial consistency in highly classifying parcels between CV and across-sample test. Since not all parcels can be expected to achieve high classification accuracies (ref), we mainly focused on the top 10% classifying parcels. Overall, our results showed that classifiers trained on single dataset samples generalized well only for certain, but not for all, test samples. In contrast, classifiers trained on the compound sample outperformed classifiers trained on single dataset samples both in terms of accuracy and consistency of accurately classifying parcels.

To evaluate generalization performance with respect to mean classification accuracies of the top 10% classifying parcels, for each pwC, we compared across-sample classification accuracies between the different test samples. Results indicate that certain datasets seem to “match” in the sense that classifiers trained on a sample from one of the datasets achieved a high accuracy when tested on the respective other one and vice versa. This was the case for HCP and 1000Brains as well as for GSP and eNKI with the former matching the results of a previous study (Weis et al., 2020). Based on the good across-sample performance of sex classifiers trained on an HCP sample on a subsample of the 1000Brains, Weis et al., (2020) suggested that parcelwise sex classification generalizes well between different samples. No additional samples from other datasets were considered in Weis et al. (2020). The present results extend the findings of the previous study by showing that good generalization performance of the HCP classifiers appears to be specific to the 1000 Brains sample. Generalization to samples from other datasets (GSP, eNKI and AOMIC) is, however, rather poor. Thus, our study demonstrates that the generalizability of pwCs trained on single dataset samples depends on the train-test data combination, which is in line with a previous study that employed sex classification based on regional homogeneity of resting state time series (Huf et al., 2014). The limited generalization performance of the pwCs trained on single dataset samples to the majority of test samples from other datasets might be attributed to the homogeneity of each single dataset training sample arising due to demographic factors such as the age range (Damoiseaux et al., 2008; Damoiseaux, 2017; Scheinost et al., 2015) as well as technical details such as fMRI acquisition parameters (Yu et al., 2018; Brown et al., 2011). Homogeneous data characteristics within each dataset will result in a homogeneity of the feature space on which ML models are trained. Such homogeneous features might lead the ML model to learn dataset specific characteristics that are predictive of the target variable, which might not translate to other test samples, resulting in inaccurate across-sample predictions (Huf et al., 2014). Thus, training ML models on a single, homogenous sample may not be ideal to achieve a good generalization performance on diverse test samples (Huf et al., 2014; Janssen et al., 2018; Belur Nagaraj et al., 2020; Di Tanna et al., 2020). In contrast, training classifiers on a combination of multiple datasets achieved significantly higher accuracies for all test samples, including the sample from a dataset which was not included in the compound training sample. The increased generalization performance might be attributable to the heterogeneity of data characteristics included in a training sample created from various datasets. This heterogeneity likely enables the model to learn patterns that do not rely on specific sample characteristics, but actually capture the underlying relationship between features and target, enabling the model to generalize better, even to data from datasets that were not included in training. Thus, training on a compound sample is preferable to training on single dataset samples (Huf et al., 2014; Chang et al., 2018; Willemink et al., 2020).

The parcelwise classification approach allowed us to investigate generalization performance not only in terms of accuracy but also with respect to the spatial distribution of accurately classifying parcels. To quantify the overlap of accurately classifying parcels between CV and across-sample test, we computed dice coefficients between within- and across sample accuracy maps at different accuracy thresholds. We observed a pattern similar to the one found for classification accuracies, with the train-test pairing of HCP and 1000Brains and GSP and eNKI, respectively, showing highest spatial consistency, relative to other combinations. Thus, also when considering spatial consistency, generalization performance depended on specific pairing of training and test datasets. For pwCs trained on single samples, training sample characteristics appeared to be the most important factor in driving generalization performance across test samples. In contrast, pwC compound achieved superior spatial consistency across all test samples, as compared to pwCs trained on single samples. Thus, the classifiers trained on the compound sample achieved both higher classification accuracies as well as more consistency in accurately classifying parcels as opposed to the classifiers trained on single dataset samples. Altogether, the high generalization performance for pwC compound can likely be attributed to the heterogeneity in the compound sample which was achieved by combining multiple samples for training. These findings match results of previous studies (Huf et al., 2014; Chang et al., 2018; Nielsen et al., 2020; Willemink et al., 2020).

Overall, the aggregation of multiple samples in pwC compound for training sex classifiers resulted in superior generalization performance. Firstly, the classification accuracies were comparable between CV and the different across-sample test classifications. Secondly, highly classifying parcels overlapped to a large degree between training and and test. The overall high generalization performance of pwC compound across all test samples could be attributed to several possible explanations: first, the compound sample is more than twice as large as compared to any of the single dataset samples. Such high sample size has been shown to be beneficial for generalization (Cui & Gong, 2018, Domingos 2012, Ishida, 2019, Yang et al. 2020). However, sample size alone is likely not sufficient to explain the high generalization performance. For instance, the eNKI sample consists of only 190 participants, but the classifiers trained on this sample achieved better generalization performance than those trained on the HCP sample, which included 878 participants. A second explanation for the good performance of pwC compound may lie in the heterogenous nature of its training sample as discussed above. Having the different samples represented within the compound sample may have allowed the classifiers to classify sex based on sample-unspecific information. Another potential explanation is that the training sample of pwC compound partially consists of data from datasets on which we evaluated the test performance. In general, training on data that is representative of the test data typically results in an increased generalization performance (Chung et al., 2018). In contrast to pwC compound, CV and across sample test performances differed considerably for pwCs trained on single dataset samples. This lack of generalization performance was especially apparent for pwC HCP which showed a rather high performance during CV in combination with the lowest generalization performance both with respect to accuracy and spatial consistency. While homogeneity of a data sample has been argued to lead to high CV classification accuracy (Huf et al., 2014), sample characteristics such as the age range were comparable between HCP and the GSP sample, with the latter outperforming HCP in generalization performance. Thus, the comparably poor performance of classifiers trained on the HCP sample may be partially attributed to sample homogeneity but also to other factors such as the differences in preprocessing pipelines. For the HCP sample, connectome extraction was based on the minimally preprocessed version of the data. The eNKI, GSP and 1000Brains samples were preprocessed using the same pipeline in SPM12, while the AOMIC sample was preprocessed using fMRIprep. Given that comparative performance evaluation of fMRI data is sensitive to preprocessing decisions (Bhagwat et al., 2021), it is likely that this difference in preprocessing may contribute to the poor generalization performance of pwC HCP when tested on the other single samples. Furthermore, the high within-sample accuracy coupled with the lack of generalization performance may also indicate an overfitting effect of pwC HCP during training (Domingos, 2012; Cui & Gong, 2018). Altogether, our results highlight the importance of a heterogenous, diverse, and representative data composition for training ML models (Gong et al., 2019; Li et al., 2022; Dhamala et al., 2023), which can be achieved by combining data from multiple sites and datasets (Nielsen et al., 2020; Willemink et al., 2020; Chang et al., 2018). By minimizing sample-specific biases, we can aim for maximizing the generalizability of ML models.

Limitations

The present results consistently demonstrated the superior generalizability of sex classifiers trained on a compound sample as compared to those trained on single dataset samples, but they come with some limitations. First of all, the high spatial consistency of pwC compound might partially be attributed to the generally higher accuracy of the across-sample predictions. Dice coefficients across the top 10% classifying parcels showed a more differentiated pattern. Here, pwC compound did not always outperform pwCs trained on single samples.

Another limitation in the present study is that, while we accounted for age as a potential confound during training of the classifiers, there might be other confounds that were not considered. For example, we did not control for structural variables such as brain size, which have been reported to influence brain functions (Batista-García-Ramó, K., & Fernández-Verdecia, C. I., 2018) and RS brain connectivity in particular (Zhang et al. ,2018). Thus, in principle, different distributions of brain size within the different samples might have influenced the present results. However, Weis et al. (2020) demonstrated that at least with their training sample, classification based on RS connectivity was not systematically influenced by brain size. Still, there might be other demographic variables which differ between samples and might influence classification accuracies (Sripada et al., 2021; Mehrabi et al., 2021; Li et al., 2022).

Another factor which has not been considered in the present analyses are fluctuating sex hormones, which have been shown to influence functional brain connectivity in RS (Weis et al., 2019; Arélin et al. 2015; Haraguchi et al. 2021). These dynamic changes in female and male connectivity patterns (Ewen & Milner, 2017; Coenjaerts et al., 2023; Kogler et al., 2016) will likely influence overall sex classification accuracies. However, unfortunately, most publicly available datasets do not provide information on hormone levels, making it impossible to consider these variations in the analyses. Future large-scale studies should include hormone levels in data acquisition, enabling model training on a combination of multiple independent datasets with well characterized phenotypes to achieve most accurate results.

Conclusion

The present results show that parcelwise sex classification models generalize best when trained on a compound sample including data with different demographic and data acquisition characteristics. Our results demonstrate that a large and heterogenous training sample including multiple datasets is best suited to achieve accurate generalization performance. This observation carries practical implications for future neuroimaging studies employing ML models for generalizable predictions.

Acknowledgements

We would like to thank Dr. Federico Raimondo and Dr. Georgios Antonopoulos and all people contributing in the "ML hour" initiative from INM-7 for the insightful discussions.

Funding:
The work was supported by the Deutsche Forschungsgemeinschaft (DFG), the National Institute of Mental Health (R01-MH074457), the Helmholtz Portfolio Theme "Supercomputing and Modeling for the Human Brain", and the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 945539 (HBP SGA3). Open access publication funded by the DFG – 491111487.

Conflicts of interest:
The authors declare no competing interests.

Data availability statement:
The datasets HCP, GSP, eNKI and AOMIC are publicly available and free to download:
https://www.humanconnectome.org/study/hcp-young-adult/data-releases
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/25833
https://openneuro.org/datasets/ds001021/versions/1.0.0
https://nilab-uva.github.io/AOMIC.github.io/
Data of the 1000Brains are available upon request from the responsible Principal Investigator (Caspers et al., 2014).
The code for preprocessing, data preparation, model training and computation of further analyses is available on Github:
https://jugit.fz-juelich.de/l.wiersch/functional_sex_classification_code
https://jugit.fz-juelich.de/f.hoffstaedter/bids_pipelines/-/tree/master/func

References

Arelin, K., Mueller, K., Barth, C., Rekkas, P. V., Kratzsch, J., Burmann, I., . . . Sacher, J. (2015). Progesterone mediates brain functional connectivity changes during the menstrual cycle-a pilot resting state MRI study. *Front Neurosci, 9*, 44. doi:10.3389/fnins.2015.00044

Arslan, A. (2018). Application of Neuroimaging in the Diagnosis and Treatment of Depression. *Understanding Depression: Volume 2. Clinical Manifestations, Diagnosis and Treatment*, 69-81.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage, 26*(3), 839–851. doi:10.1016/j.neuroimage.2005.02.018

Batista-Garcia-Ramo, K., & Fernandez-Verdecia, C. I. (2018). What We Know About the Brain Structure-Function Relationship. *Behav Sci (Basel), 8*(4). doi:10.3390/bs8040039

Belur Nagaraj, S., Pena, M. J., Ju, W., Heerspink, H. L., & BEAt-DKD Consortium. (2020). Machine-learning–based early prediction of end-stage renal disease in patients with diabetic kidney disease using clinical trials data. *Diabetes, Obesity and Metabolism, 22*(12), 2479-2486.

Betzel, R. F., Byrge, L., He, Y., Goni, J., Zuo, X. N., & Sporns, O. (2014). Changes in structural and functional connectivity among resting-state networks across the human lifespan. *NeuroImage, 102 Pt 2*, 345-357. doi:10.1016/j.neuroimage.2014.07.067

Bhagwat, N., Barry, A., Dickie, E. W., Brown, S. T., Devenyi, G. A., Hatano, K., . . . Poline, J. B. (2021). Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses. *Gigascience, 10*(1). doi:10.1093/gigascience/giaa155

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, 144-152.

Brown, G. G., Mathalon, D. H., Stern, H., Ford, J., Mueller, B., Greve, D. N., . . . Function Biomedical Informatics Research, N. (2011). Multisite reliability of cognitive BOLD data. *NeuroImage, 54*(3), 2163-2175. doi:10.1016/j.neuroimage.2010.09.076

Buch, V. H., Ahmed, I., & Maruthappu, M. (2018). Artificial intelligence in medicine: current trends and future possibilities. *Br J Gen Pract, 68*(668), 143-144. doi:10.3399/bjgp18X695213

Caspers, S., Moebus, S., Lux, S., Pundt, N., Schutz, H., Muhleisen, T. W., . . . Amunts, K. (2014). Studying variability in human brain aging in a population-based German cohort-rationale and design of 1000BRAINS. *Front Aging Neurosci, 6*, 149. doi:10.3389/fnagi.2014.00149

Chang, K., Balachandar, N., Lam, C., Yi, D., Brown, J., Beers, A., . . . Kalpathy-Cramer, J. (2018). Distributed deep learning networks among institutions for medical imaging. *J Am Med Inform Assoc, 25*(8), 945-954. doi:10.1093/jamia/ocy017

Chen, J., Patil, K. R., Weis, S., Sim, K., Nickl-Jockschat, T., Zhou, J., . . . Outcome Survey, I. (2020). Neurobiological Divergence of the Positive and Negative Schizophrenia Subtypes Identified on a New Factor Structure of Psychopathology Using Non-negative Factorization: An International Machine Learning Study. *Biol Psychiatry, 87*(3), 282-293. doi:10.1016/j.biopsych.2019.08.031

Chung, Y., Haas, P. J., Upfal, E., & Kraska, T. (2018). Unknown examples & machine learning model generalization. *arXiv preprint arXiv:1808.08294.*

Coenjaerts, M., Adrovic, B., Trimborn, I., Philipsen, A., Hurlemann, R., & Scheele, D. (2023). Effects of exogenous oxytocin and estradiol on resting-state functional connectivity in women and men. *Sci Rep, 13*(1), 3113. doi:10.1038/s41598-023-29754-y

Cohen, A. D., Chen, Z., Parker Jones, O., Niu, C., & Wang, Y. (2020). Regression-based machine-learning approaches to predict task activation using resting-state fMRI. *Hum Brain Mapp, 41*(3), 815-826. doi:10.1002/hbm.24841

Cui, Z., & Gong, G. (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *NeuroImage, 178*, 622-637. doi:10.1016/j.neuroimage.2018.06.001

Damoiseaux, J. S. (2017). Effects of aging on functional and structural brain connectivity. *NeuroImage, 160*, 32-40. doi:10.1016/j.neuroimage.2017.01.077

Damoiseaux, J. S., Beckmann, C. F., Arigita, E. J., Barkhof, F., Scheltens, P., Stam, C. J., . . . Rombouts, S. A. (2008). Reduced resting-state brain activity in the "default network" in normal aging. *Cereb Cortex, 18*(8), 1856-1864. doi:10.1093/cercor/bhm207

Dhamala, E., Yeo, B. T. T., & Holmes, A. J. (2023). One Size Does Not Fit All: Methodological Considerations for Brain-Based Predictive Modeling in Psychiatry. *Biol Psychiatry, 93*(8), 717-728. doi:10.1016/j.biopsych.2022.09.024

Di Tanna, G. L., Wirtz, H., Burrows, K. L., & Globe, G. (2020). Correction: Evaluating risk prediction models for adults with heart failure: A systematic literature review. *PloS one, 15*(7), e0235970. doi:10.1371/journal.pone.0235970

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*(3).

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87.

Esteban, O., Ciric, R., Finc, K., Blair, R. W., Markiewicz, C. J., Moodie, C. A., . . . Gorgolewski, K. J. (2020). Analysis of task-based functional MRI data preprocessed with fMRIPrep. *Nat Protoc, 15*(7), 2186-2202. doi:10.1038/s41596-020-0327-3

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., . . . Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat Methods, 16*(1), 111-116. doi:10.1038/s41592-018-0235-4

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., . . . Jiang, T. (2016). The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cereb Cortex, 26*(8), 3508-3526. doi:10.1093/cercor/bhw157

Genon, S., Eickhoff, S. B., & Kharabian, S. (2022). Linking interindividual variability in brain structure to behaviour. *Nat Rev Neurosci, 23*(5), 307-318. doi:10.1038/s41583-022-00584-7

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., . . . Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage, 80,* 105–124. doi:10.1016/j.neuroimage.2013.04.127

Gong, Z., Zhong, P., & Hu, W. (2019). Diversity in machine learning. *Ieee Access, 7,* 64323-64350.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front Neuroinform, 5*, 13. doi:10.3389/fninf.2011.00013

Haraguchi, R., Hoshi, H., Ichikawa, S., Hanyu, M., Nakamura, K., Fukasawa, K., ... & Shigihara, Y. (2021). The menstrual cycle alters resting-state cortical activity: a magnetoencephalography study. *Frontiers in human neuroscience*.

Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat Rev Neurosci, 7*(7), 523-534. doi:10.1038/nrn1931

Holmes, A. J., Hollinshead, M. O., O'Keefe, T. M., Petrov, V. I., Fariello, G. R., Wald, L. L., . . . Buckner, R. L. (2015). Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. *Sci Data, 2*, 150031. doi:10.1038/sdata.2015.31

Holmes, C. J., Hoge, R., Collins, L., Woods, R., Toga, A. W.,, & Evans, A. C. (1998). Enhancement of MR images using registration for signal averaging. *Journal of computer assisted tomography, 22*(2), 324–333.

Huf, W., Kalcher, K., Boubela, R. N., Rath, G., Vecsei, A., Filzmoser, P., & Moser, E. (2014). On the generalizability of resting-state fMRI machine learning classifiers. *Frontiers in human neuroscience, 8*, 502.

Huntenberg, J. M. (2014). Evaluating nonlinear coregistration of BOLD EPI and T1w images. *(Doctoral dissertation, Freie Universität Berlin)*.

Ishida, E. E. (2019). Machine learning and the future of supernova cosmology. *Nature Astronomy, 3*(8), 680-682.

Jansma, J. M., Rutten, G. J., Ramsey, L. E., Snijders, T. J., Bizzi, A., Rosengarth, K., . . . Ramsey, N. F. (2020). Correction to: Automatic identification of atypical clinical fMRI results. *Neuroradiology, 62*(12), 1723. doi:10.1007/s00234-020-02565-y

Janssen, R. J., Mourao-Miranda, J., & Schnack, H. G. (2018). Making Individual Prognoses in Psychiatry Using Neuroimaging and Machine Learning. *Biol Psychiatry Cogn Neurosci Neuroimaging, 3*(9), 798-808. doi:10.1016/j.bpsc.2018.04.004

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage, 17*(2), 825-841. doi:10.1016/s1053-8119(02)91132-8

Jollans, L., Boyle, R., Artiges, E., Banaschewski, T., Desrivieres, S., Grigis, A., . . . Whelan, R. (2019). Quantifying performance of machine learning methods for neuroimaging data. *NeuroImage, 199*, 351-365. doi:10.1016/j.neuroimage.2019.05.082

Kogler, L., Muller, V. I., Seidel, E. M., Boubela, R., Kalcher, K., Moser, E., . . . Derntl, B. (2016). Sex differences in the functional connectivity of the amygdalae in association with cortisol. *NeuroImage, 134*, 410-423. doi:10.1016/j.neuroimage.2016.03.064

Kohoutova, L., Heo, J., Cha, S., Lee, S., Moon, T., Wager, T. D., & Woo, C. W. (2020). Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nat Protoc, 15*(4), 1399-1435. doi:10.1038/s41596-019-0289-5

Li, J., Bzdok, D., Chen, J., Tam, A., Ooi, L. Q. R., Holmes, A. J., . . . Genon, S. (2022). Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci Adv, 8*(11), eabj1812. doi:10.1126/sciadv.abj1812

McEwen, B. S., & Milner, T. A. (2017). Understanding the broad influence of sex hormones and sex differences in the brain. *J Neurosci Res, 95*(1-2), 24-39. doi:10.1002/jnr.23809

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR), 54*(6), 1-35.

Menon, S. S., & Krishnamurthy, K. (2019). A Comparison of Static and Dynamic Functional Connectivities for Identifying Subjects and Biological Sex Using Intrinsic Individual Brain Connectivity. *Sci Rep, 9*(1), 5729. doi:10.1038/s41598-019-42090-4

More, S., Antonopoulos, G., Hoffstaedter, F., Caspers, J., Eickhoff, S. B., Patil, K. R., & Alzheimer's Disease Neuroimaging, I. (2023). Brain-age prediction: A systematic comparison of machine learning workflows. *NeuroImage, 270*, 119947. doi:10.1016/j.neuroimage.2023.119947

Nielsen, A. N., Barch, D. M., Petersen, S. E., Schlaggar, B. L., & Greene, D. J. (2020). Machine Learning With Neuroimaging: Evaluating Its Applications in Psychiatry. *Biol Psychiatry Cogn Neurosci Neuroimaging, 5*(8), 791-798. doi:10.1016/j.bpsc.2019.11.007

Nooner, K. B., Colcombe, S. J., Tobe, R. H., Mennes, M., Benedict, M. M., Moreno, A. L., . . . Milham, M. P. (2012). The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. *Frontiers in neuroscience, 6*, 152. doi:10.3389/fnins.2012.00152

Nostro, A. D., Müller, V. I., Varikuti, D. P., Pläschke, R. N., Hoffstaedter, F., Langner, R., . . . Eickhoff, S. B. (2018). Predicting personality from network-based resting-state functional connectivity. *Brain structure & function, 223*(6), 2699–2719. doi:10.1007/s00429-018-1651-z

Plaschke, R. N., Patil, K. R., Cieslik, E. C., Nostro, A. D., Varikuti, D. P., Plachti, A., . . . Eickhoff, S. B. (2020). Age differences in predicting working memory performance from network-based functional connectivity. *Cortex, 132*, 441-459. doi:10.1016/j.cortex.2020.08.012

Rafi, M., & Shaikh, M. S. (2013). A comparison of SVM and RVM for Document Classification. *arXiv preprint arXiv:1301.2785*.

Rutten, G. J., & Ramsey, N. F. (2010). The role of functional magnetic resonance imaging in brain surgery. *Neurosurg Focus, 28*(2), E4. doi:10.3171/2009.12.FOCUS09251

Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L., & Smith, S. M. (2014). Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage, 90*, 449–468. doi:10.1016/j.neuroimage.2013.11.046

Sanford, N., Ge, R., Antoniades, M., Modabbernia, A., Haas, S. S., Whalley, H. C., . . . Frangou, S. (2022). Sex differences in predictors and regional patterns of brain age gap estimates. *Hum Brain Mapp, 43*(15), 4689-4698. doi:10.1002/hbm.25983

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., . . . Wolf, D. H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage, 64*, 240-256. doi:10.1016/j.neuroimage.2012.08.052

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X. N., Holmes, A. J., . . . Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral

Cortex from Intrinsic Functional Connectivity MRI. *Cereb Cortex, 28*(9), 3095-3114. doi:10.1093/cercor/bhx179

Scheinost, D., Finn, E. S., Tokoglu, F., Shen, X., Papademetris, X., Hampson, M., & Constable, R. T. (2015). Sex differences in normal age trajectories of functional brain networks. *Human brain mapping, 36*(4), 1524–1535. doi:10.1002/hbm.22720

Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., . . . Glasser, M. F. (2013). Resting-state fMRI in the Human Connectome Project. *NeuroImage, 80*, 144–168. doi:10.1016/j.neuroimage.2013.05.039

Smith, S. M., Nichols, T. E., Vidaurre, D., Winkler, A. M., Behrens, T. E., Glasser, M. F., . . . Miller, K. L. (2015). A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nat Neurosci, 18*(11), 1565-1567. doi:10.1038/nn.4125

Smith, S. M., Vidaurre, D., Beckmann, C. F., Glasser, M. F., Jenkinson, M., Miller, K. L., . . . Van Essen, D. C. (2013). Functional connectomics from resting-state fMRI. *Trends Cogn Sci, 17*(12), 666-682. doi:10.1016/j.tics.2013.09.016

Snoek, L., van der Miesen, M. M., Beemsterboer, T., van der Leij, A., Eigenhuis, A., & Steven Scholte, H. (2021). The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses. *Sci Data, 8*(1), 85. doi:10.1038/s41597-021-00870-6

Sorensen, T. A. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skar., 5*, 1-34.

Sripada, C., Angstadt, M., Taxali, A., Clark, D. A., Greathouse, T., Rutherford, S., . . . Heitzeg, M. (2021). Brain-wide functional connectivity patterns support general cognitive ability and mediate effects of socioeconomic status in youth. *Transl Psychiatry, 11*(1), 571. doi:10.1038/s41398-021-01704-0

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., & Consortium, W. U.-M. H. (2013). The WU-Minn Human Connectome Project: an overview. *NeuroImage, 80*, 62-79. doi:10.1016/j.neuroimage.2013.05.041

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., . . . Consortium, W. U.-M. H. (2012). The Human Connectome Project: a data acquisition perspective. *NeuroImage, 62*(4), 2222-2231. doi:10.1016/j.neuroimage.2012.02.018

Vapnik, V. (1998). Statistical learning theory Wiley. *New York, 1*(624), 2.

Varikuti, D. P., Genon, S., Sotiras, A., Schwender, H., Hoffstaedter, F., Patil, K. R., . . . Eickhoff, S. B. (2018). Evaluation of non-negative matrix factorization of grey matter in age prediction. *NeuroImage, 173*, 394-410. doi:10.1016/j.neuroimage.2018.03.007

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage, 180*(Pt A), 68-77. doi:10.1016/j.neuroimage.2017.06.061

Varoquaux, G., & Thirion, B. (2014). How machine learning is shaping cognitive neuroimaging. *Gigascience, 3*(1), 1-7.

Wang, S., Peterson, D. J., Gatenby, J. C., Li, W., Grabowski, T. J., & Madhyastha, T. M. (2017). Evaluation of Field Map and Nonlinear Registration Methods for Correction of Susceptibility Artifacts in Diffusion MRI. *Front Neuroinform, 11*, 17. doi:10.3389/fninf.2017.00017

Weis, S., Hodgetts, S., & Hausmann, M. (2019). Sex differences and menstrual cycle effects in cognitive and sensory resting state networks. *Brain Cogn, 131*, 66-73. doi:10.1016/j.bandc.2017.09.003

Weis, S., Patil, K. R., Hoffstaedter, F., Nostro, A., Yeo, B. T. T., & Eickhoff, S. B. (2020). Sex Classification by Resting State Brain Connectivity. *Cereb Cortex, 30*(2), 824-835. doi:10.1093/cercor/bhz129

Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., . . . Lungren, M. P. (2020). Preparing Medical Imaging Data for Machine Learning. *Radiology, 295*(1), 4-15. doi:10.1148/radiol.2020192224

Yang, F., Wanik, D. W., Cerrai, D., Bhuiyan, M. A. E., & Anagnostou, E. N. (2020). Quantifying uncertainty in machine learning-based power outage prediction model training: A tool for sustainable storm restoration. *Sustainability, 12*(4), 1525.

Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., . . . Sheline, Y. I. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Hum Brain Mapp, 39*(11), 4213-4227. doi:10.1002/hbm.24241

Zhang, C., Dougherty, C. C., Baum, S. A., White, T., & Michael, A. M. (2018). Functional connectivity predicts gender: Evidence for gender differences in resting brain connectivity. *Human brain mapping, 39*(4), 1765–1776. doi:10.1002/hbm.23950

Zhang, Z., Li, G., Xu, Y., & Tang, X. (2021). Application of Artificial Intelligence in the MRI Classification Task of Human Brain Neurological and Psychiatric Diseases: A Scoping Review. *Diagnostics (Basel), 11*(8). doi:10.3390/diagnostics11081402