Received: 29 August 2023 | Revised: 19 March 2024 | Accepted: 2 April 2024

DOI: 10.1002/hbm.26683

RESEARCH ARTICLE

WILEY

Sex classification from functional brain connectivity: Generalization to multiple datasets

Lisa Wiersch[1,2] | Patrick Friedrich[1,2] | Sami Hamdan[1,2] | Vera Komeyer[1,2,3] | Felix Hoffstaedter[1,2] | Kaustubh R. Patil[1,2] | Simon B. Eickhoff[1,2] | 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

[3]Department of Biology, Faculty of Mathematics and Natural Sciences, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Correspondence
Susanne Weis, Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany.
Email: s.weis@fz-juelich.de

Funding information
Deutsche Forschungsgemeinschaft (DFG), Grant/Award Numbers: 491111487, 431549029; National Institute of Mental Health, Grant/Award Number: R01-MH074457; the Helmholtz Portfolio Theme "Supercomputing and Modeling for the Human Brain"; European Union's Horizon 2020 Research and Innovation Programme, Grant/Award Number: 945539

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 compound samples of two different sizes. Generalization performance was quantified in terms of mean across-sample classification accuracy and spatial consistency of accurately classifying parcels. Our results indicate that the generalization performance of parcelwise classifiers (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 pwCs trained on the compound samples 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 both a large sample size and a heterogeneous data composition of a training sample have a central role in achieving generalizable results.

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

1 | 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 (Buch et al., 2018; Jollans et al., 2019; Kohoutova et al., 2020; Varoquaux, 2018). Successful applications of ML approaches include the decoding of mental states (Haynes & Rees, 2006), classification of mental disorders (Chen et al., 2020; Zhang et al., 2021), as well as the prediction of demographic and behavioral phenotypes (More et al., 2023; Nostro et al., 2018; Pläschke et al., 2020; Smith et al., 2015; Varikuti et al., 2018).

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

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2024 The Authors. *Human Brain Mapping* published by Wiley Periodicals LLC.

(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 et al., 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; Smith, Vidaurre, et al., 2013; Zhang et al., 2018), 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 (Cui & Gong, 2018; Smith et al., 2015) 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 & Ramsey, 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. Due to the variability across different datasets and sites, a ML model trained on a compound of such data is more likely to capture the shared biological variability in all datasets while disregarding the variability resulting from differences between the datasets. This distinction supports models focusing solely on the biological variability independent of specific dataset characteristics. Hence, such models are less likely to overfit and more likely to generalize to new data. 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 (Chang et al., 2018; Nielsen et al., 2020; Willemink et al., 2020). 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, we aimed to evaluate the generalization performance of multiple sets of sex classification models derived from different training samples. The different training samples were created from four different datasets with varying demographic characteristics. In addition, sex classifiers were trained on compound samples combining data from all datasets to obtain training samples with heterogeneous sample characteristics. Both compound samples comprise the same ratios of datasets, sex and age distributions, but differ in sample size to additionally assess the effect of training sample size. 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 six training samples, resulting in six 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 pwCs 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 pwC trained on the compound sample with a smaller sample size should outperform pwCs trained on single samples due to the heterogeneous data composition, while the compound sample with a higher sample size should achieve the overall best generalization performance (Chang et al., 2018; Cui & Gong, 2018; Dhamala et al., 2023; Domingos, 2012; Nielsen et al., 2020; Willemink et al., 2020).

2 | MATERIALS AND METHODS

2.1 | Data

We employed RS 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 Superstructure 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 ("compound854") was constructed by combining subsamples of the HCP, GSP, eNKI and 1000Brains samples, with a mean age of 40.05 (range: 20–85), resulting in a sample size of 854 subjects. This sample size is equal to the GSP sample, but larger than the eNKI and lower than the HCP and 1000Brains samples, therefore representing an intermediate sample size compared to the other data samples. Another sixth sample ("compound2190") was constructed by combining 75% of the HCP, GSP, eNKI and 1000Brains samples resulting in a sample size of 2190 subjects in total. The compound 2190 sample comprised a mean age of 40.10 years (range: 20–85 years). Thus, both compound samples display a large difference in sample size but ratios of dataset representation, sex and age distribution have been maintained. This allows us to evaluate the influence of data composition compared to the sample size of a training sample on the generalization performance of sex classification models.

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.

2.2 | Data acquisition

2.2.1 | HCP

The RS fMRI data of the HCP dataset were acquired on a Siemens Skyra 3 T MRI scanner with multiband echo-planar imaging with a duration of 873 s and the following parameters: 72 slices; voxel size, 2 × 2 × 2 mm^3; field of view (FOV), 208 × 180 mm^2; matrix, 104 × 90; TR, 720 ms; TE, 33 ms; flip angle, 52° (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, Beckmann, et al., 2013).

2.2.2 | GSP

RS data were acquired on a 3 T Tim Trio Scanner with a duration of 372 s and the following parameters: 47 slices; voxel size, 3 × 3 × 3 mm^3; FOV, 216 mm; TR, 3 s; TE, 30 ms; flip angle, 85°. 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).

2.2.3 | eNKI

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

2.2.4 | 1000Brains

Subjects were scanned for 660 s on a Siemens TRIO 3 T MRI scanner with the following parameters: 36 slices; voxel size, 3.1 × 3.1 × 3.1 mm^3; FOV, 200 × 200 mm^2; matrix, 64 × 64, TR = 2.2 s; TE = 30 ms; flip angle, 90°. 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).

2.2.5 | AOMIC

The AOMIC dataset includes two subsamples, PIOP1 and PIOP2, comprising data of healthy university students scanned on a Philips 3 T 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 × 3 × 3 mm^3 and a matrix size of 80 × 80. While PIOP1 was acquired for 360 s with multi-slice acceleration, 480 volumes and a 0.75 TR, PIOP2 was acquired for 480 s without multi-slice acceleration, 240 volumes and a 2 s TR (further details in https://www.nature.com/articles/s41597-021-00870-6/tables/10).

2.3 | Data preprocessing

2.3.1 | HCP

RS data from the 'HCP S1200 Release' analyzed here was fully preprocessed and denoised via the Connectome Workbench software. In short, data were corrected for spatial distortions, head motion, B_0 distortions and were registered to the T1-weighted structural image (Smith, Beckmann, 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 2 mm isotropic 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 2 mm Full width half maximum (FWHM) surface smoothing (Glasser et al., 2013; Smith, Beckmann, 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.

2.3.2 | GSP, eNKI, 1000Brains

RS data of the GSP, eNKI and 1000Brains samples were preprocessed in the same way. Initially, FSL was used for the removal of noise and motion artifacts by applying the FIX-denoising procedure (Jenkinson et al., 2012; Salimi-Khorshidi et al., 2014) using the appropriate pre-trained dataset for noise classification. As FIX does not include normalization to MNI space, 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 mean EPI image was spatially normalized to the MNI152 template (Holmes et al., 1998) using the "unified segmentation" approach (Ashburner & Friston, 2005) and the resulting deformation was applied to the FIX-denoised images and resampled to 2 mm^3.

2.3.3 | AOMIC

Fully preprocessed data was used provided via OpenNeuro, where it was 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 (Huntenberg, 2014; Wang et al., 2017) with six 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).

2.4 | 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.

2.5 | Parcelwise sex classification

Sex classification models were trained based on the individual multivariate RSFC profile of each parcel. Specifically, the connectivity values between each parcel and the 435 remaining parcels were used as features to train a sex classification model per parcel, resulting in a set of 436 pwC (Weis et al., 2020). Since each model provides one final accuracy value, one pwC provides an accuracy map covering the entire brain. Training sex classification models based on the connectivity profile of each parcel allows for a reduction of the feature dimensionality for each model (1 × 436) as compared to training one model based on the overall connectivity profile (436 × 436). Furthermore, using this parcelwise approach allows us to identify the highest classifying brain regions. In the following steps, we evaluated generalization performance in terms of classification accuracies and spatial consistency of highly classifying parcels across the entire brain.

All models were built using support vector machine (SVM) classifiers. SVM is a supervised ML method that separates the data into distinct classes with the widest possible gap between these classes (Boser et al., 1992; Rafi & Shaikh, 2013; Vapnik, 1998; Zhang et al., 2021). Based on its operational principles regarding a supervised binary classification task and successful applications in previous sex classification studies (Flint et al., 2020; Weis et al., 2020; Wiersch et al., 2023), SVM is a suitable method for the present task. SVM models were built in Julearn (Hamdan et al., 2023; https://juaml.github.io/julearn/main/index.html) including a hyperparameter search nested within a 10–fold CV with five repetitions. The parameter search included choice of kernel (linear vs. radial basis function (rbf) kernel) as well as the hyperparameters gamma and C, which is used to set the strength of regularization (https://scikit-learn.org/stable/auto_examples/svm/plot_svm_scale_c.html). The SVM algorithm used in the present study incorporates a squared L2 regularization. The regularization parameter controls the trade-off between the model fit to the training data and generalizable predictions beyond the training data in order to avoid overfitting and to optimize model performance and generalizability (https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html).

Confounding effects of age were regressed out in a CV-consistent manner by removing age-related variance before training the classifiers. By estimating confound regression models only for training subsets and applying them to training and test sets, leakage of information from test to training data within the CV-process can be avoided (More et al., 2021). 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 a cross-sample classification, single dataset pwCs were tested on the respective other three samples, while pwC compound 854 and pwC compound 2190 were tested on the remaining 25% of the HCP ($n = 220$, mean age: 29.68, age range: 22–36), GSP ($n = 214$, mean age: 22–72, age range: 21–31), eNKI ($n = 48$, mean age: 47.52, age range: 20–75) and 1000Brains ($n = 250$, mean age: 52.08, age range: 22–80) sample. 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 six pwCs was assessed on the AOMIC sample. All reported accuracies are balanced accuracies.

2.6 | Statistical analyses

2.6.1 | Across-sample classification accuracy

To statistically compare the classification accuracies of pwCs across the different test samples, we employed independent *t*-tests between the different across-sample accuracies over the respectively 10% highest classifying parcels. Additional analyses using all 436 parcels are reported in the supplements (Table S3 and below).

Significance levels were Bonferroni-corrected according to the number of dependent tests (15 dependent tests for comparing across-sample accuracies of all six pwCs on the AOMIC test sample, 10 dependent tests for comparing the across-sample accuracy of both compound pwCs for the five test samples and for comparing pwC performances against each other for each of the five test samples; six dependent tests for all other comparisons).

2.6.2 | 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).

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

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.

3 | RESULTS

The generalization performance of pwCs trained on each of the single dataset samples (HCP, GSP, eNKI, & 1000Brains) and on both compound samples 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 to determine whether pwCs trained on compound samples exhibit more generalizable results in contrast to pwCs trained on single samples.

3.1 | 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 3 (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.

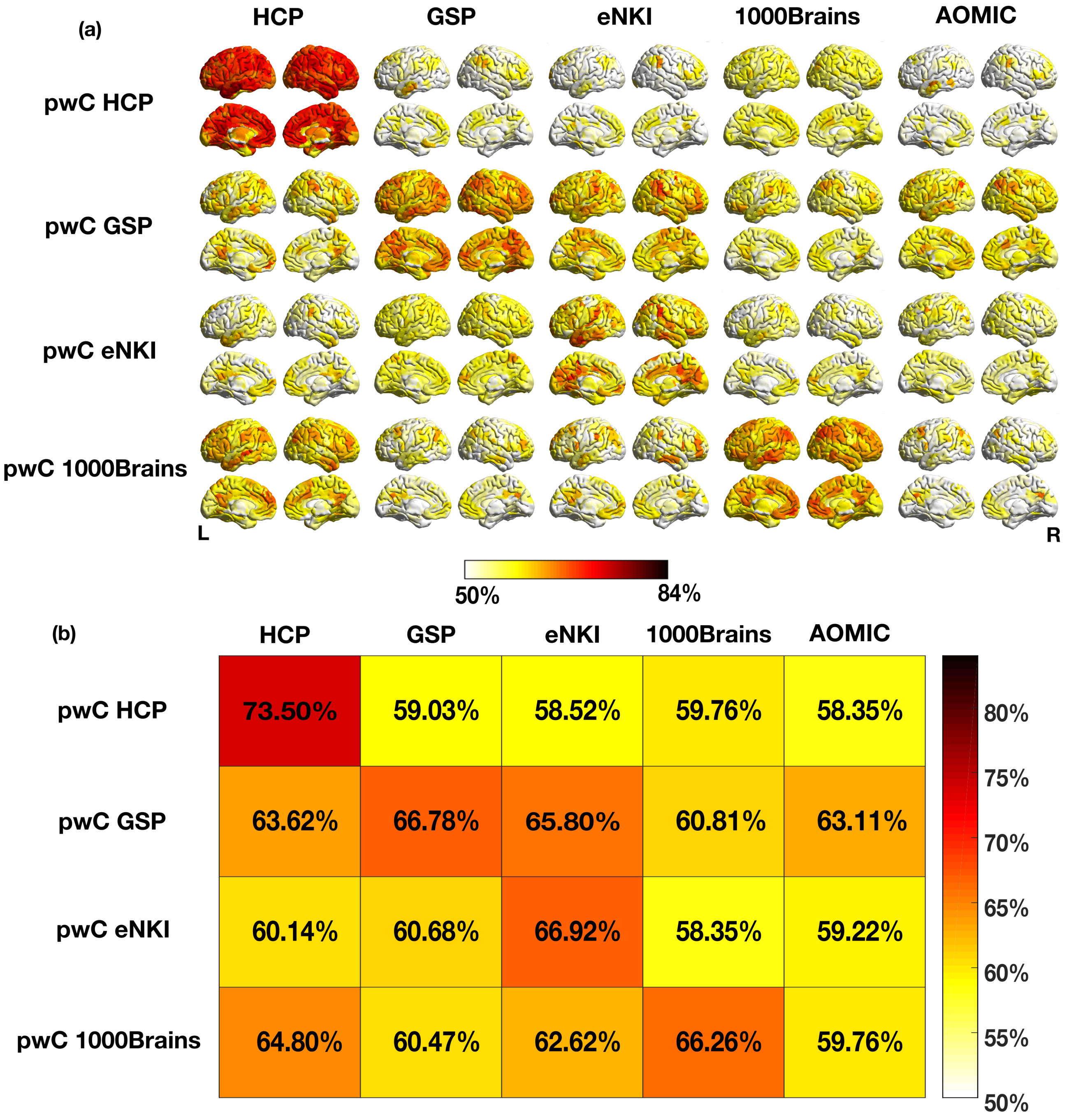

FIGURE 1 Accuracy maps and tile plots of mean accuracies of top 10% classifying parcels for parcelwise classifiers (pwCs) trained on single samples. (a) Spatial distribution of parcelwise sex classification accuracies across the brain. Within-sample accuracies are depicted on and across-sample accuracies 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 cross-validation (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

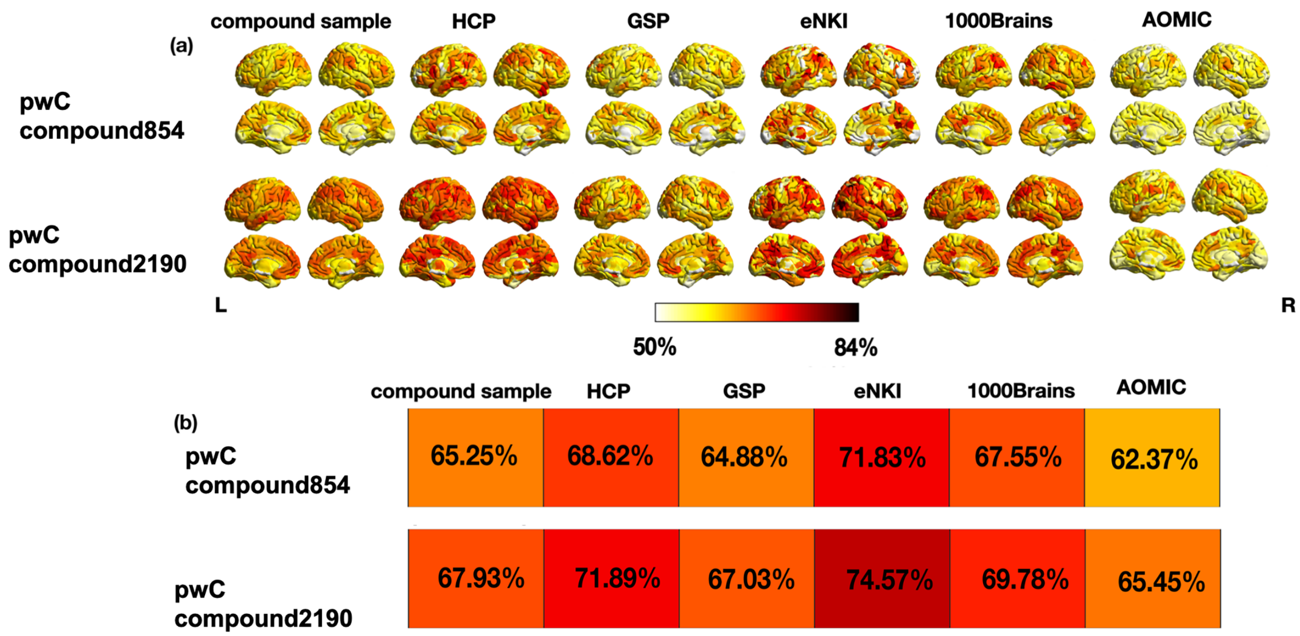

FIGURE 2 Accuracy maps and tile plots of mean accuracies of top 10% classifying parcels for parcelwise classifiers (pwC) compound 854 and pwC compound 2190. (a) Spatial distribution of parcelwise sex classification accuracies across the brain. Only parcels with an accuracy of 0.5 or higher are displayed. (b) Mean accuracies averaged across the top 10% classifying parcels for the respective cross-validation (CV)- (first column) and across-sample predictions.

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 compound854 achieved a mean within-sample accuracy of 65.3% for the top 10% classifying parcels, while mean across-sample accuracies of the highest classifying parcels ranged between 62.4% (pwC compound854 tested on AOMIC) and 71.8% (pwC compound854 tested on eNKI, further details in Table S1 and Figure 2, Figure S2). PwC compound2190 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 compound2190 tested on AOMIC) and 74.6% (pwC compound2190 tested on eNKI, details in Table S1 and Figure 2, Figure S2).

Contrasting the top 10% classifying parcels in the accuracy maps of pwC compound854 and pwC compound2190 displayed peaks in accuracies for the eNKI test sample (71.8% and 74.6%) resulting in significantly higher accuracies than for the remaining test samples, respectively (Figure 2 and Table S2). We also contrasted how the six pwCs performed on each test sample by employing independent t-tests: pwC compound 854 outperformed all pwCs trained on single samples for all test samples, except for the AOMIC test sample, where pwC GSP achieved higher accuracies within the best 10% classifying parcels (Table S2). PwC compound 2190 outperformed all other pwCs for the HCP, GSP, eNKI and AOMIC test sample with regards to the top 10% classifying parcels in each across-sample prediction (Figure 2). Details for all statistical comparisons are shown in Table S2.

3.2 | Consistency of correctly classifying parcels

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 3. Regarding spatial consistency within a given pwC (horizontal comparison in Figure 3), 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). Spatial consistency of pwC compound854 ranged between 0.2865–0.3221 for the HCP, GSP, eNKI and 1000Brains sample and achieved 0.2546 for the AOMIC sample. PwC compound2190 demonstrated a 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 Figure 3) pwC compound854 demonstrated higher spatial consistency than single sample

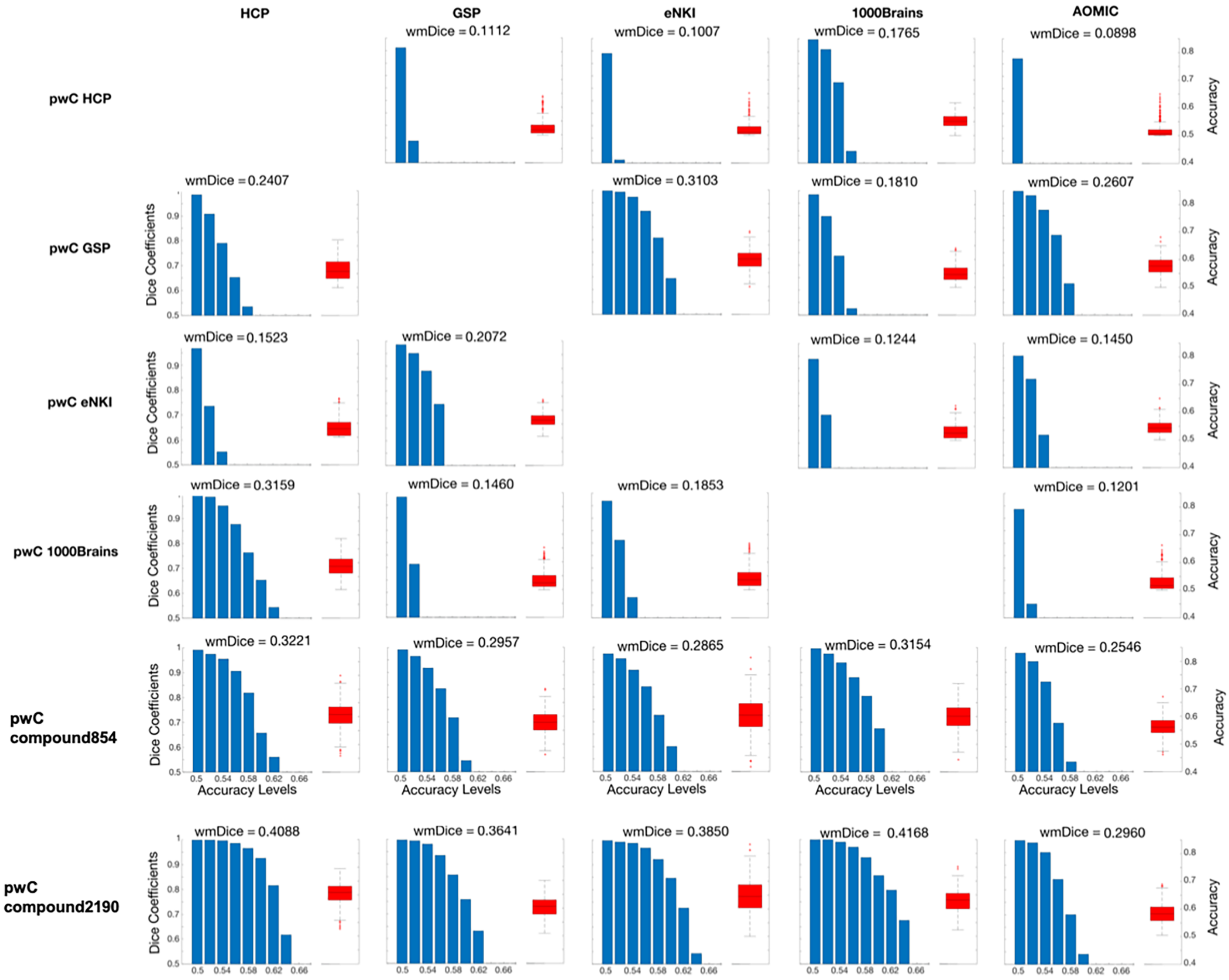

FIGURE 3 Spatial consistency of all parcelwise classifiers (pwCs). For each combination of training (rows) and test sample (columns), the right side of each subplot (red boxplot) depicts the distribution of accuracies across all parcels (right y-axis). The left side of each subplot (blue barplot) shows the dice coefficients (left y-axis), representing the overlap of accuracy maps between cross-validation (CV) and test predictions 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.

pwCs for the HCP, GSP and 1000Brains test samples and pwC compound2190 demonstrated higher spatial consistency than the other six pwCs. Dice coefficients for the top 10% classifying parcels are reported in Figure S3.

4 | 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 testing. Since not all parcels are expected to achieve high classification accuracies (Weis et al., 2020; Zhang et al., 2018), 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 test samples. In contrast, classifiers trained on the compound samples tend to outperform 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 1000Brains 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 RS 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, 2017; Damoiseaux et al., 2008; Scheinost et al., 2015) as well as technical details such as fMRI acquisition parameters (Brown et al., 2011; Yu et al., 2018). 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 (Belur Nagaraj et al., 2020; Di Tanna et al., 2020; Huf et al., 2014; Janssen et al., 2018). In contrast, training classifiers on a combination of multiple datasets (pwC compound854 and pwC compound2190) achieved significantly higher accuracies for all test samples, including the sample from a dataset which was not included in the compound training sample. We contrasted performances of both pwCs trained on compound samples to evaluate potential sample size effects. Here, pwC compound854 demonstrated higher accuracies and spatial consistency in the majority of across-sample predictions compared to single sample pwCs, but did not outperform pwC compound2190. These results suggest that the sample size of the training sample is an important factor in determining the generalization performance of ML analyses. These results align with the findings of several other studies highlighting the importance of the sample size in ensuring accurate ML results (Cui & Gong, 2018; Dhamala et al., 2023; Domingos, 2012; Ishida, 2019; Yang et al., 2020). However, pwC compound854 still predominantly demonstrated a higher generalization performance compared to single sample pwCs with a similar or even higher sample size. Thus, it is evident that the composition of a training sample is crucial in ensuring generalizable ML results, as reported by previous studies (Chang et al., 2018; Huf et al., 2014; Willemink et al., 2020). While a high sample size is beneficial to assure reliable and accurate ML predictions (Dhamala et al., 2023), the heterogeneity and representativeness of a composite sample led to significantly better results than single sample pwCs with a higher sample size in the present ML analyses. Thus, the high generalization performance of both compound samples is not only attributable to the sample size but also 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. Therefore, the heterogeneity of a composite training sample is essential for generalizable ML outcomes and may also serve to minimize sample-specific biases (Li et al., 2022). Thus, training on a compound sample comprising the variability of multiple datasets is preferable to training on single dataset samples in order to achieve high generalization performances (Chang et al., 2018; Huf et al., 2014; Willemink et al., 2020).

While undesirable sources of variability, e.g. due to scanner differences, may be accounted for by using data harmonization (Fortin et al., 2017; Yu et al., 2018), in the present study we intentionally refrained from using harmonization techniques. Here, we evaluated the generalization performances of differently trained pwCs in order to determine which may generalize best to unseen data. Harmonization techniques such as ComBat are not suitable for this purpose because they require a sufficient amount of data from each sample and site (Orlhac et al., 2022).

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 testing, 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 the 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 compound854 achieved superior spatial consistency in most test samples and pwC compound2190 in all test samples, as compared to pwCs trained on single samples. Thus, the classifiers trained on the compound samples 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 compound854 and pwC compound2190 can likely be attributed to the data heterogeneity in the respective training samples which was achieved by combining multiple samples for training. These findings match results of previous studies (Chang et al., 2018; Huf et al., 2014; Nielsen et al., 2020; Willemink et al., 2020).

Overall, the aggregation of multiple samples in pwC compound854 and pwC compound2190 for training sex classifiers resulted in superior generalization performance compared to pwCs trained on single samples. 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 test. The overall high generalization performance of pwC compound2190 across all test samples could be attributed to several possible explanations: first, the compound2190 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. In addition, analyses with pwC compound854 also demonstrated a superior generalization performance with respect to classification accuracies as well as spatial consistency compared to single sample pwCs, despite the smaller sample size. A second explanation for the good performance of both compound pwCs may lie in the heterogeneous 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 samples of pwC compound854 and pwC compound2190 partially consist 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). Here, both training samples for the compound pwCs composed data from four different datasets. Although each dataset had a different sample size and thus a different share in the respective compound training sample, the model applications to the eNKI test sample showed highest accuracies for the best 10% classifying parcels. This result stems from few parcels classifying at a high level for the eNKI test data (up to 83%), resulting in such a high mean accuracy for the top 10% parcels (Figure 2). Furthermore, the mean accuracy averaged across all 436 parcels confirms that there are only few parcels responsible for the high accuracy in the top 10% parcels, as the eNKI dataset did not exhibit the overall highest mean accuracy across all parcels.

In contrast to both compound pwCs, 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 FIX denoised preprocessed version of the data. The eNKI, GSP and 1000Brains samples were preprocessed using the same pipeline in FSL/SPM12 also including FIX-denoising, while the AOMIC sample was preprocessed using fMRIprep without FIX. 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 (Cui & Gong, 2018; Domingos, 2012).

The present study, however, does not primarily aim to build a classifier attaining highest sex classification accuracies but rather to evaluate the impact of the training sample in ML models, particularly the size and composition of the training sample.

Altogether, our results highlight the importance of the sample size and also a heterogeneous, diverse, and representative data composition for training ML models (Cui & Gong, 2018; Dhamala et al., 2023; Domingos, 2012; Gong et al., 2019; Li et al., 2022), which can be achieved by combining data from multiple sites and datasets (Chang et al., 2018; Nielsen et al., 2020; Willemink et al., 2020). By minimizing sample-specific biases, we can aim for maximizing the generalizability of ML models.

4.1 | Limitations

The present results consistently demonstrated the superior generalizability of sex classifiers trained on compound samples as compared to those trained on single dataset samples, but they come with some limitations. First of all, the high spatial consistency of pwC compound2190 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 compound2190 did not always outperform pwCs trained on single samples. Overall, the predominantly higher generalization performance of pwC compound2190 can be attributed to the sample size and sample composition of its training sample. However, an additional systematic study would be required to determine the exact degree to which each factor contributes to high generalization performance.

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-Garcia-Ramo & Fernandez-Verdecia, 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 (Li et al., 2022; Mehrabi et al., 2021; Sripada et al., 2021).

A further limitation of the present study is the potential impact of different preprocessing approaches which may affect the outcomes in ML analyses. In neuroimaging data, there can be various sources of noise and artifacts. Prior to data analysis, it is necessary to preprocess the data to mitigate these issues and enhance the data quality. However, the impact of preprocessing steps on the outcomes of fMRI analyses has been well documented. For instance, conceptually similar preprocessing packages such as AFNI, FSL, or SPM can produce differences in fMRI results (Bowring et al., 2019). Differences on the level of preprocessing steps may also produce dissimilarities (Carp, 2012). Even differences in the order of preprocessing steps can lead to differences in the graph theoretical outcomes derived from RS functional connectivity (Gargouri et al., 2018). Thus, it is plausible that discrepancies in preprocessing pipelines may lead to differences in classification outcomes. Indeed, one study that compared ML results for patient and healthy control classification across different preprocessing pipelines indicated differences in the classification accuracy (Vergara et al., 2017). Overall, while different preprocessing approaches may lead to differences in the fMRI and ML results, in the present study these differences represent an additional source of variance that may occur when using data of various datasets. Despite various potential sources of variance within the training samples of the compound pwCs, pwC compound854 and pwC compound2190 demonstrate a comparatively good performance compared to the single sample pwCs. While it is reasonable to anticipate that aligned preprocessing approaches may improve predictions; however, conducting a systematic evaluation on the effect of preprocessing pipelines is beyond the scope of the present study and remains an important open question for future research.

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 (Arélin et al., 2015; Haraguchi et al., 2021; Weis et al., 2019). These dynamic changes in female and male connectivity patterns (Coenjaerts et al., 2023; Kogler et al., 2016; McEwen & Milner, 2017) 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.

5 | 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 heterogeneous 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.

ACKNOWLEDGMENTS

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. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

The work was supported by the Deutsche Forschungsgemeinschaft (DFG - German Research Foundation) – Project-ID 431549029 - Collaborative Research Centre CRC1451 on motor performance project B05, 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.

CONFLICT OF INTEREST STATEMENT

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.

ORCID

Lisa Wiersch https://orcid.org/0000-0001-8006-8678

REFERENCES

Arélin, K., Mueller, K., Barth, C., Rekkas, P. V., Kratzsch, J., Burmann, I., Villringer, A., & Sacher, J. (2015). Progesterone mediates brain functional connectivity changes during the menstrual cycle-a pilot resting state MRI study. *Frontiers in Neuroscience*, *9*, 44. https://doi.org/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*, *2*, 69–81.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, *26*(3), 839–851. https://doi.org/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. *Behavioral Sciences*, *8*(4), 1–14. https://doi.org/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. https://doi.org/10.1016/j.neuroimage.2014.07.067

Bhagwat, N., Barry, A., Dickie, E. W., Brown, S. T., Devenyi, G. A., Hatano, K., DuPre, E., Dagher, A., Chakravarty, M., Greenwood, C. M. T., Misic, B., Kennedy, D. N., & Poline, J. B. (2021). Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses. *GigaScience*, *10*(1), 1–13. https://doi.org/10.1093/gigascience/giaa155

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144–152). Association for Computing Machinery.

Bowring, A., Maumet, C., & Nichols, T. E. (2019). Exploring the impact of analysis software on task fMRI results. *Human Brain Mapping*, *40*(11), 3362–3384.

Brown, G. G., Mathalon, D. H., Stern, H., Ford, J., Mueller, B., Greve, D. N., McCarthy, G., Voyvodic, J., Glover, G., Diaz, M., Yetter, E., Ozyurt, I. B., Jorgensen, K. W., Wible, C. G., Turner, J. A., Thompson, W. K., Potkin, S. G., & Function Biomedical Informatics Research Network. (2011). Multisite reliability of cognitive BOLD data. *NeuroImage*, *54*(3), 2163–2175. https://doi.org/10.1016/j.neuroimage.2010.09.076

Buch, V. H., Ahmed, I., & Maruthappu, M. (2018). Artificial intelligence in medicine: current trends and future possibilities. *The British Journal of General Practice*, *68*(668), 143–144. https://doi.org/10.3399/bjgp18X695213

Carp, J. (2012). On the plurality of (methodological) worlds: Estimating the analytic flexibility of FMRI experiments. *Frontiers in Neuroscience*, *6*, 149.

Caspers, S., Moebus, S., Lux, S., Pundt, N., Schutz, H., Muhleisen, T. W., Gras, V., Eickhoff, S. B., Romanzetti, S., Stöcker, T., Stirnberg, R., Kirlangic, M. E., Minnerop, M., Pieperhoff, P., Mödder, U., Das, S., Evans, A. C., Jöckel, K. H., Erbel, R., … Amunts, K. (2014). Studying variability in human brain aging in a population-based German cohort-rationale and design of 1000BRAINS. *Frontiers in Aging Neuroscience*, *6*, 149. https://doi.org/10.3389/fnagi.2014.00149

Chang, K., Balachandar, N., Lam, C., Yi, D., Brown, J., Beers, A., Rosen, B., Rubin, D. L., & Kalpathy-Cramer, J. (2018). Distributed deep learning networks among institutions for medical imaging. *Journal of the American Medical Informatics Association*, *25*(8), 945–954. https://doi.org/10.1093/jamia/ocy017

Chen, J., Patil, K. R., Weis, S., Sim, K., Nickl-Jockschat, T., Zhou, J., Aleman, A., Sommer, I. E., Liemburg, E. J., Hoffstaedter, F., Habel, U., Derntl, B., Liu, X., Fischer, J. M., Kogler, L., Regenbogen, C., Diwadkar, V. A., Stanley, J. A., Riedl, V., … Pharmacotherapy Monitoring and Outcome Survey (PHAMOUS) Investigators. (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. *Biological Psychiatry*, *87*(3), 282–293. https://doi.org/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. *Scientific Reports*, *13*(1), 3113. https://doi.org/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. *Human Brain Mapping*, *41*(3), 815–826. https://doi.org/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. https://doi.org/10.1016/j.neuroimage.2018.06.001

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

Damoiseaux, J. S., Beckmann, C. F., Arigita, E. J., Barkhof, F., Scheltens, P., Stam, C. J., Smith, S. M., & Rombouts, S. A. (2008). Reduced resting-state brain activity in the "default network" in normal aging. *Cerebral Cortex*, *18*(8), 1856–1864. https://doi.org/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. *Biological Psychiatry*, *93*(8), 717–728. https://doi.org/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. https://doi.org/10.1371/journal.pone.0235970

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302.

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., Kent, J. D., Goncalves, M., DuPre, E., Gomez, D. E. P., Ye, Z., Salo, T., Valabregue, R., Amlien, I. K., Liem, F., Jacoby, N., Stojić, H., Cieslak, M., Urchs, S., … Gorgolewski, K. J. (2020). Analysis of task-based functional MRI data preprocessed with fMRIPrep. *Nature Protocols*, *15*(7), 2186–2202. https://doi.org/10.1038/s41596-020-0327-3

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox, P. T., Eickhoff, S. B., Yu, C., & Jiang, T. (2016). The human Brainnetome atlas: A new brain atlas based on connectional architecture. *Cerebral Cortex*, *26*(8), 3508–3526. https://doi.org/10.1093/cercor/bhw157

Flint, C., Förster, K., Koser, S. A., Konrad, C., Zwitserlood, P., Berger, K., Hermesdorf, M., Kircher, T., Nenadic, I., Krug, A., Baune, B. T., Dohm, K., Redlich, R., Opel, N., Arolt, V., Hahn, T., Jiang, X., Dannlowski, U., & Grotegerd, D. (2020). Biological sex classification with structural MRI data shows increased misclassification in transgender women. *Neuropsychopharmacology*, *45*(10), 1758–1765. https://doi.org/10.1038/s41386-020-0666-3

Fortin, J. P., Parker, D., Tunc, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., Schultz, R. T., Verma, R., & Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, *161*, 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047

Gargouri, F., Kallel, F., Delphine, S., Ben Hamida, A., Lehéricy, S., & Valabregue, R. (2018). The influence of preprocessing steps on graph theory measures derived from resting state fMRI. *Frontiers in Computational Neuroscience*, *12*, 8.

Genon, S., Eickhoff, S. B., & Kharabian, S. (2022). Linking interindividual variability in brain structure to behaviour. *Nature Reviews. Neuroscience*, *23*(5), 307–318. https://doi.org/10.1038/s41583-022-00584-7

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., Jenkinson, M., & WU-Minn HCP Consortium. (2013). The minimal preprocessing pipelines for the human connectome project. *NeuroImage*, *80*, 105–124. https://doi.org/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. *Frontiers in Neuroinformatics*, *5*, 13. https://doi.org/10.3389/fninf.2011.00013

Hamdan, S., More, S., Sasse, L., Komeyer, V., Patil, K. R., & Raimondo, F. (2023). Julearn: An easy-to-use library for leakage-free evaluation and inspection of ML models. *arXiv Preprint arXiv:2310.12568*.

Haraguchi, R., Hoshi, H., Ichikawa, S., Hanyu, M., Nakamura, K., Fukasawa, K., Poza, J., Rodríguez-González, V., Gómez, C., & Shigihara, Y. (2021). The menstrual cycle alters resting-state cortical activity: A magnetoencephalography study. *Frontiers in Human Neuroscience*, *15*, 652789.

Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews. Neuroscience*, *7*(7), 523–534. https://doi.org/10.1038/nrn1931

Holmes, A. J., Hollinshead, M. O., O'Keefe, T. M., Petrov, V. I., Fariello, G. R., Wald, L. L., Fischl, B., Rosen, B. R., Mair, R. W., Roffman, J. L., Smoller, J. W., & Buckner, R. L. (2015). Brain genomics Superstruct project initial data release with structural, functional, and behavioral measures. *Scientific Data*, *2*, 150031. https://doi.org/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., Dodoo-Schittko, F., Hattingen, E., de la Peña, M. J., von Campe, G., Jehna, M., & Ramsey, N. F. (2020). Correction to: Automatic identification of atypical clinical fMRI results. *Neuroradiology*, *62*(12), 1723. https://doi.org/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. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(9), 798–808. https://doi.org/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. https://doi.org/10.1016/s1053-8119(02)91132-8

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *Neuroimage*, *62*(2), 782–790.

Jollans, L., Boyle, R., Artiges, E., Banaschewski, T., Desrivieres, S., Grigis, A., Martinot, J. L., Paus, T., Smolka, M. N., Walter, H., Schumann, G., Garavan, H., & Whelan, R. (2019). Quantifying performance of machine learning methods for neuroimaging data. *NeuroImage*, *199*, 351–365. https://doi.org/10.1016/j.neuroimage.2019.05.082

Kogler, L., Muller, V. I., Seidel, E. M., Boubela, R., Kalcher, K., Moser, E., Habel, U., Gur, R. C., Eickhoff, S. B., & Derntl, B. (2016). Sex differences in the functional connectivity of the amygdalae in association with cortisol. *NeuroImage*, *134*, 410–423. https://doi.org/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. *Nature Protocols*, *15*(4), 1399–1435. https://doi.org/10.1038/s41596-019-0289-5

Li, J., Bzdok, D., Chen, J., Tam, A., Ooi, L. Q. R., Holmes, A. J., Ge, T., Patil, K. R., Jabbi, M., Eickhoff, S. B., Yeo, B. T. T., & Genon, S. (2022). Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Science Advances*, *8*(11), eabj1812. https://doi.org/10.1126/sciadv.abj1812

McEwen, B. S., & Milner, T. A. (2017). Understanding the broad influence of sex hormones and sex differences in the brain. *Journal of Neuroscience Research*, *95*(1–2), 24–39. https://doi.org/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. *Scientific Reports*, *9*(1), 5729. https://doi.org/10.1038/s41598-019-42090-4

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

More, S., Eickhoff, S. B., Caspers, J., & Patil, K. R. (2021). Confound removal and normalization in practice: A neuroimaging based sex prediction case study. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 3–18.

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. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(8), 791–798. https://doi.org/10.1016/j.bpsc.2019.11.007

Nooner, K. B., Colcombe, S. J., Tobe, R. H., Mennes, M., Benedict, M. M., Moreno, A. L., Panek, L. J., Brown, S., Zavitz, S. T., Li, Q., Sikka, S., Gutman, D., Bangaru, S., Schlachter, R. T., Kamiel, S. M., Anwar, A. R., Hinz, C. M., Kaplan, M. S., Rachlin, A. B., … Milham, M. P. (2012). The NKI-Rockland sample: A model for accelerating the pace of discovery science in psychiatry. *Frontiers in Neuroscience*, *6*, 152. https://doi.org/10.3389/fnins.2012.00152

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

Orlhac, F., Eertink, J. J., Cottereau, A. S., Zijlstra, J. M., Thieblemont, C., Meignan, M., Boellaard, R., & Buvat, I. (2022). A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *Journal of Nuclear Medicine*, *63*(2), 172–179.

Pläschke, R. N., Patil, K. R., Cieslik, E. C., Nostro, A. D., Varikuti, D. P., Plachti, A., Lösche, P., Hoffstaedter, F., Kalenscher, T., Langner, R., & Eickhoff, S. B. (2020). Age differences in predicting working memory performance from network-based functional connectivity. *Cortex*, *132*, 441–459. https://doi.org/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. *Neurosurgical Focus*, *28*(2), E4. https://doi.org/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. https://doi.org/10.1016/j.neuroimage.2013.11.046

Sanford, N., Ge, R., Antoniades, M., Modabbernia, A., Haas, S. S., Whalley, H. C., Galea, L., Popescu, S. G., Cole, J. H., & Frangou, S. (2022). Sex differences in predictors and regional patterns of brain age gap estimates. *Human Brain Mapping*, *43*(15), 4689–4698. https://doi.org/10.1002/hbm.25983

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., Gur, R. 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. https://doi.org/10.1016/j.neuroimage.2012.08.052

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X. N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-global Parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, *28*(9), 3095–3114. https://doi.org/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. https://doi.org/10.1002/hbm.22720

Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., Ugurbil, K., Barch, D. M., Van Essen, D. C., & Glasser, M. F. (2013). Resting-state fMRI in the human connectome project. *NeuroImage*, *80*, 144–168. https://doi.org/10.1016/j.neuroimage.2013.05.039

Smith, S. M., Nichols, T. E., Vidaurre, D., Winkler, A. M., Behrens, T. E., Glasser, M. F., Ugurbil, K., Barch, D. M., Van Essen, D. C., & Miller, K. L. (2015). A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience*, *18*(11), 1565–1567. https://doi.org/10.1038/nn.4125

Smith, S. M., Vidaurre, D., Beckmann, C. F., Glasser, M. F., Jenkinson, M., Miller, K. L., Nichols, T. E., Robinson, E. C., Salimi-Khorshidi, G., Woolrich, M. W., Barch, D. M., Uğurbil, K., & Van Essen, D. C. (2013). Functional connectomics from resting-state fMRI. *Trends in Cognitive Sciences*, *17*(12), 666–682. https://doi.org/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. *Scientific Data*, *8*(1), 85. https://doi.org/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. *I kommission hos E. Munksgaard*, *5*, 1–34.

Sripada, C., Angstadt, M., Taxali, A., Clark, D. A., Greathouse, T., Rutherford, S., Dickens, J. R., Shedden, K., Gard, A. M., Hyde, L. W., Weigard, A., & Heitzeg, M. (2021). Brain-wide functional connectivity patterns support general cognitive ability and mediate effects of socioeconomic status in youth. *Translational Psychiatry*, *11*(1), 571. https://doi.org/10.1038/s41398-021-01704-0

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., & WU-Minn HCP Consortium. (2013). The WU-Minn human connectome project: An overview. *NeuroImage*, *80*, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., Della Penna, S., Feinberg, D., Glasser, M. F., Harel, N., Heath, A. C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., … WU-Minn HCP Consortium. (2012). The human connectome project: A data acquisition perspective. *NeuroImage*, *62*(4), 2222–2231. https://doi.org/10.1016/j.neuroimage.2012.02.018

Vapnik, V. (1998). *Statistical learning theory* (p. 2). Wiley.

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

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

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

Vergara, V. M., Mayer, A. R., Damaraju, E., & Calhoun, V. D. (2017). The effect of preprocessing in dynamic functional network connectivity used to classify mild traumatic brain injury. *Brain and Behavior*, *7*(10), e00809.

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. *Frontiers in Neuroinformatics*, *11*, 17. https://doi.org/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 and Cognition*, *131*, 66–73. https://doi.org/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. *Cerebral Cortex*, *30*(2), 824–835. https://doi.org/10.1093/cercor/bhz129

Wiersch, L., Hamdan, S., Hoffstaedter, F., Votinov, M., Habel, U., Clemens, B., … Weis, S. (2023). Accurate sex prediction of cisgender and transgender individuals without brain size bias. *Scientific Reports*, *13*(1), 13868.

Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L. R., Summers, R. M., Rubin, D. L., & Lungren, M. P. (2020). Preparing medical imaging data for machine learning. *Radiology*, *295*(1), 4–15. https://doi.org/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., Trivedi, M. H., Weissman, M. M., Shinohara, R. T., & Sheline, Y. I. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Human Brain Mapping*, *39*(11), 4213–4227. https://doi.org/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. https://doi.org/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), 1402. https://doi.org/10.3390/diagnostics11081402

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wiersch, L., Friedrich, P., Hamdan, S., Komeyer, V., Hoffstaedter, F., Patil, K. R., Eickhoff, S. B., & Weis, S. (2024). Sex classification from functional brain connectivity: Generalization to multiple datasets. *Human Brain Mapping*, *45*(6), e26683. https://doi.org/10.1002/hbm.26683