Title: Machine Learning And MRI-Based Diagnostic Models For ADHD: Are We There Yet?

Authors: Yanli Zhang-James[1], Martine Hoogman[2,3,4], Barbara Franke[2,3,4], and Stephen V Faraone[1,5*]

Affiliations:
Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, New York
Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands
Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands
Department of Psychiatry, Radboud University Medical Center, Nijmegen, The Netherlands
Department of Neuroscience and Physiology, SUNY Upstate Medical University, Syracuse, New York

Corresponding author:
Stephen V. Faraone, Ph. D.
SUNY Upstate Medical University
750 E Adam St,
Syracuse, NY, 13210
315-464-3113, 315-849-1839 (fax)
sfaraone@childpsychresearch.org

Short running title: MRI Classifiers for ADHD

Declaration of interest

Drs Zhang-James and Hoogman declare no conflict of interest.
Dr. Franke has received educational speaking fees from Medice.
In the past year, Dr. Faraone received income, potential income, travel expenses continuing education support and/or research support from Takeda, OnDosis, Tris, Otsuka, Arbor, Ironshore, Rhodes, Akili Interactive Labs, Sunovion, Supernus and Genomind. With his institution, he has US patent US20130217707 A1 for the use of sodium-hydrogen exchange inhibitors in the treatment of ADHD. He also receives royalties from books published by Guilford Press: Straight Talk about Your Child's Mental Health, Oxford University Press: Schizophrenia: The Facts and Elsevier: ADHD: Non-Pharmacologic Interventions. He is Program Director of www.adhdinadults.com.

*Corresponding author

Abstract

Machine learning (ML) has been applied to develop magnetic resonance imaging (MRI)-based diagnostic classifiers for attention-deficit/hyperactivity disorder (ADHD). This systematic review examines this literature to clarify its clinical significance and to assess the implications of the various analytic methods applied. We found that, although most of studies reported the classification accuracies, they varied in choice of MRI modalities, ML models, cross-validation and testing methods, and sample sizes. We found that the accuracies of cross-validation methods inflated the performance estimation compared with those of a held-out test, compromising the model generalizability. Test accuracies have increased with publication year but were not associated with training sample sizes. Improved test accuracy over time was likely due to the use of better ML methods along with strategies to deal with data imbalances. Ultimately, large multi-modal imaging datasets, and potentially the combination with other types of data, like cognitive data and/or genetics, will be essential to achieve the goal of developing clinically useful imaging classification tools for ADHD in the future.

Key words: attention deficit hyperactivity disorder; biomarkers; classification; machine learning; MRI

Introduction

Clinicians diagnose ADHD by evaluating symptoms of hyperactivity, impulsivity, inattention, and impaired functioning across settings. The diagnosis of ADHD shows considerable levels of concurrent and predictive validity in its clinical features, course, neurobiology, and treatment response (Faraone, 2005; Faraone & Biederman, 2000). Nevertheless, concerns about diagnostic accuracy persist. Some suggest that the current method of diagnosing ADHD is too subjective and leads to over-diagnosing ADHD in the community (Bruchmuller, Margraf, & Schneider, 2012; Visser et al., 2014). Psychiatric diagnoses have been called "subjective" because they rely on clinician evaluation of responses from patients, parents, and/or informants. Other studies have raised concerns about the under-diagnosis of ADHD (Ginsberg, Quintero, Anand, Casillas, & Upadhyaya, 2014; The Express Scripts Lab, 2014), especially in girls and women, which suggests biases in applying the current diagnostic algorithm. Another issue is the misdiagnosis of ADHD as being another disorder. When this occurs, patients may be exposed to unnecessary treatments and will continue to struggle with the many impairments associated with ADHD. Those who have ADHD and are not diagnosed with the disorder will continue to have impaired functioning leading to increased risks for other health and social problems (Dalsgaard, Ostergaard, Leckman, Mortensen, & Pedersen, 2015; Franke et al., 2018; Lambert & Hartsough, 1998; Lichtenstein et al., 2012; Reiersen & Todorov, 2013).

In response to these concerns, researchers have sought to develop objective measures to diagnose ADHD or to monitor the course of ADHD symptoms during treatment. Much research has examined peripheral biochemical markers in differentiating ADHD and control patients, such as (Norepinephrine (NE), 3-Methoxy-4-hydroxyphenylethylene glycol (MHPG), monoamine oxidase (MAO), zinc, and cortisol (Faraone, Bonvicini, & Scassellati, 2014; Scassellati, Bonvicini, Faraone, & Gennarelli, 2012). NE, MHPG, MAO, b-phenylethylamine, and cortisol are also somewhat predictive of response to ADHD medications. Meta-analysis also shows that peripheral measures of oxidative stress differ between ADHD and control participants (Joseph, Zhang-James, Perl, & Faraone, 2015). Electroencephalographic (EEG) (Snyder, Rugino, Hornig, & Stein, 2015), actigraphic (Dane, Schachar, & Tannock, 2000), and eye vergence measurements (Sole Puig et al., 2015), as well as interactive gaming behaviour (Faraone et al., 2016) were also examined as ADHD biomarkers. Neuropsychological tests (Ritsner, 2009), particularly continuous performance tests (CPTs) (e.g. Corkum & Siegel, 1993; Homack & Riccio, 2006; Riccio & Reynolds, 2001) have been evaluated in many studies. In recent years, genetic markers in the form of polygenic risk scores also have shown some predictive ability of diagnosis and prognosis of ADHD (Demontis et al., 2019; Hamshere et al., 2013; Riglin et al., 2016).Many of these prior studies show group differences but do not present diagnostic accuracy statistics. A clinically useful biomarker should have at least 80% sensitivity and 80% specificity. They should also be reliable, reproducible, inexpensive, non-invasive, easy to use, and confirmed by at least two independent studies. These criteria were defined by work of the task force on biological markers by the World Federation of ADHD (Thome et al., 2012). None of the measures examined by them met these criteria for clinical utility (Thome et al., 2012).

Prior structural MRI (sMRI) studies have consistently reported alterations in frontal, parietotemporal, cingulate, cerebellum, basal ganglia, and corpus callosum regions (Castellanos et al., 2002; Mackie et al., 2007; Seidman, Valera, & Makris, 2005; Seidman et al., 2006; Shaw et al., 2014; Shaw et al., 2006; Valera, Faraone, Murray, & Seidman, 2007). Studies of the largest ADHD sMRI dataset from the Enhancing Neuro Imaging Genetics Through Meta-Analysis (ENIGMA) consortium’s ADHD Working Group reported the significant volumetric reductions in intracranial volume, amygdala, caudate nucleus, nucleus accumbens, hippocampus, and cortical surface areas from many regions in children with ADHD (Hoogman et al., 2017; Hoogman et al., 2019). These regions have also be implicated in functional MRI (fMRI) studies showing altered brain connectivity and activation in the fronto-striatal, frontoparietal and fronto-temporal-parietal circuits, as well dorsal anterior cingulate cortex in ADHD brains (Dickstein, Bannon, Castellanos, & Milham, 2006; Smith, Taylor, Brammer, Toone, & Rubia, 2006; Tian et al., 2006). Studies have also examined the developmental trajectories of these anatomical and functional alterations across the lifespan finding initial delays that are followed by apparent normalization (Castellanos et al., 2002; Shaw et al., 2014; Shaw et al., 2006).

These findings encouraged efforts to develop objective diagnostic tools for ADHD using MRI data. Early studies used standard statistical methods such as discriminant analysis with very small sample sizes (Semrud-Clikeman et al., 1996; Zhu et al., 2008; Zhu et al., 2005). For example, a discriminant analysis reported by Semrud-Clikeman et al. (1996) included 10 participants in each of three diagnostic groups: developmental dyslexia, ADHD or control. Zhu and colleagues' discriminant analysis classifier assessed 9 ADHD and 11 typically developing boys. Although high predictive accuracies were reported in these studies (85~ 87%), it is difficulty to evaluate how well those models would generalize given the small samples and lack of replication samples.

The ADHD-200 Global Competition (Consortium, 2012) challenged researchers to develop an MRI-based diagnostic classifier for ADHD. It provided a dataset of 776 children, adolescents and young adults (7-21 years old, 63% healthy controls, 37% ADHD) from eight sites. Fifty teams from around the world joined the competition with 21 final submissions. Machine learning models predominated. Due to the large sample size, the consortium was able to set aside a test set that was not used for model selection and development. The competition results were judged by the performance on the test set only. This contrasts with previous studies with small sample sizes, where a held-out test set was not available. The ADHD-200 winning team used an ensemble model which achieved a 61% accuracy with 21% sensitivity and 94% specificity using both structural and resting state-fMRI data along with the demographic predictors (Eloyan et al., 2012). The accuracy, although considerably lower than previously reported high accuracies, was one of the first in an independent, held-out test set. Despite the modest accuracy, the ADHD-200 competition re-kindled enthusiasm for developing imaging-based diagnostic classifiers for ADHD. The publicly available dataset has become the main data source driving the machine learning model development for ADHD. Since the competition in 2012, we have seen a steady increase in the number of publications applying machine learning classifiers to ADHD. Thirty-one additional published studies have used either the whole or part of the ADHD-200 dataset (Supplementary Figure 1 and Table 1).

This systematic review examines the prior literature applying ML to MRI data in ADHD to clarify the clinical significance of findings and to assess the implications of the various analytic methods applied. We discuss the progress made over the years as well as lessons and methodological issues that we learned from this body of work. We hope to provide a roadmap for future studies that aim to overcome these issues and achieve clinically useful models for diagnosing ADHD.

Methods

A literature search on MRI-based diagnostic classifiers for ADHD using key words (“ADHD” AND “MRI” AND (“Machine learning” OR “Classi*)) and examining their references identified 49 studies in total (up to June 20, 2020, Pubmed, Embase and Google). Supplementary Figure 2 shows the article selection procedure in a PRISMA diagram. The eligible studies applied statistical or machine learning classifiers using MRI data to differentiate participants with ADHD from controls. Table 1 lists the selected studies along with the performance of their best models. If a study dealt with multi-class classification, for example, having ADHD, ASD and control groups, only the two-class classification accuracies involving ADHD vs the control groups were examined in this review. We used percent correct (accuracy) to compare results across studies because it was available for most of the papers. Studies that met the classifier criteria but did not report an accuracy statistic or other metrics that can be used to compute accuracies, were not included in our quantitative analysis. If a study reported multiple models, only the model which had the highest accuracies was included in Table 1.

We extracted and examined study characteristics, including machine learning model types, MRI data modality, cross-validation and testing methods, training sample size, training set class ratio (the ratios of ADHD vs Control participants' numbers), data source, dataset age and sex compositions and publication years, etc. We grouped machine learning models to three categories: support vector machine (SVM), convolutional neural networks (CNNs), and others. We assigned studies with a training set class ratio between 0.4 ~0.6 as "balanced" (i.e., nearly equal), and those with higher or lower ratios as "unbalanced". Six studies used various methods to balance demographic differences between the ADHD and control groups. These were assigned as "balanced", even if their original class ratio was outside of the balanced range (Deshpande, Wang, Rangaprakash, & Wilamowski, 2015; Fair et al., 2012; Ghiassian, Greiner, Jin, & Brown, 2016; M. N. Qureshi, B. Min, H. J. Jo, & B. Lee, 2016; Riaz, Asad, Alonso, & Slabaugh, 2018; Wang, Jiao, Tang, Wang, & Lu, 2013). We reported the age and sex groups, as well as the minimum and maximum age range of the dataset. For the ADHD-200 samples, the overall age range was used if a specific subset was used but age information was not provided. Minimum and maximum values of age were derived for studies that reported mean and standard deviation of the ages.

We also classified studies based on the methods they used to evaluate model performance and generalizability. Two methods were used. The held-out test set method evaluates model performance on data that were set aside, i.e., they were not used during model estimation and training. Because this method requires a large sample, many studies resort to cross-validation

(CV) method to assess model performance. CV methods randomly re-sample examples to be set aside during model fitting. The most commonly used versions are the leave-one-out CV (LOOCV) (Deshpande et al., 2015; Fair et al., 2012; Hart et al., 2014; Iannaccone et al., 2015; Peng, Lin, Zhang, & Wang, 2013) and K-Fold-CV (where K is often = 10, 5, or 2) (Brown et al., 2012; Dai, Wang, Hua, & He, 2012; Du, Wang, Jie, & Zhang, 2016; M. N. I. Qureshi, B. Min, H. J. Jo, & B. Lee, 2016). For example, in 10-fold CV, the original dataset is partitioned into 10 equal sub-samples or "folds". For each iteration of model estimation nine of the subsamples are used to estimate model parameters and the left-out fold is used to estimate model accuracy. The left-out fold changes from iteration to iteration. For LOOCV, one sample is left out for testing while all the others are used for training or model fitting. In either situation, the process is repeated until all samples have been used in both the training and test sets. The CV accuracy is estimated by averaging over all iterations of CV accuracies. Although CV samples were not used during the model training/fitting at each iteration, they are, nevertheless, used as training examples in other iterations.

Our main objective was to understand how study features influenced model accuracy. We used likelihood-ratio (LR) test assisted variable selection in combination with multivariate linear regression to quantitatively evaluate if these features predicted model accuracy. The variable selection algorithm implemented in STATA16's *gvselect* command computes both the Akaike's (1974) information criterion (AIC) and Schwarz's (1978) Bayesian information criterion (BIC) (StataCorp, 2019). We performed the variable selection and linear regression modeling for all the studies combined, as well as separately for the K-Fold-CV, LOOCV, and held-out test groups. Training sample size was primarily examined as a continuous variable. However, we also classified sample sizes as small (<300) or large (>300) to compare the variability of their accuracy estimates using Levene's robust test statistic (Levene, 1960). In addition to the quantitative analysis, we also qualitatively reviewed the relevant study characteristics if a quantitative analysis was not possible.

Results

Among all the studies included, over half the studies (N=27, 55%) reported only CV results without a held-out test set. Forty-three percent (N=21) used a held-out test sample to evaluate classifier performance. All but one of the 21 studies used the ADHD-200 samples. Among the studies that reported held-out test results, six also reported CV results. Figure 1 shows that the 16 studies using K-Fold-CV and the 17 studies using LOOCV reported, on average, higher accuracies than studies using held-out tests ($F_{(2, 47)} = 25.3$, $p < 0.001$).

Accuracy estimates increased in later publication years ($F_{(1, 47)}=13.68$, p=0.0006, Figure 2). This effect was driven by the studies using held-out test sets ($F_{(1, 21)}=11.0$, p=0.003, Figure 2). There was no significant change of reported accuracies over the years for studies using the K-Fold-CV or LOOCV methods (Figure 2).

Training sample size significantly predicted accuracies in the K-Fold-CV group ($F_{(1, 15)} = 7.8$, p = 0.01, Figure 3 Left). Studies with large samples had lower mean accuracies than studies with small studies (71% vs 79% mean accuracies, $F_{(1, 15)} = 6.5$, p = 0.02). The sample size effect was not significant for the LOOCV and held-out test groups, (Figure 3 Middle and Right), either as a continuous or categorical variable. The accuracy results from small studies were more variable than those from large studies in the held-out test group (Levene's robust test statistic $W0_{(1, 27)} = 6.68$, p = 0.015). The variance differences between large and small samples were not statistically significant for either the K-Fold-CV or LOOCV.

Twenty-four studies (55%) used a training dataset that had severely imbalanced classes. Six of those studies applied data balancing methods to compensate for the class imbalance and are grouped as balanced studies. Class-balanced studies reported higher accuracies for both the K-Fold-CV ($F_{(1, 15)}$ = 6.5, p = 0.02) and LOOCV ($F_{(1, 17)}$ = 22.2, p = 0.0002, Supplementary Figure 3A). However, the balanced studies in the K-Fold-CV group were all small studies (Supplementary Figure 3B); we could not differentiate whether the higher accuracy was due to the negative relationship with sample size or the benefit of data balance. The higher accuracies in the balanced LOOCV group was not related to sample size. No statistical difference was found for either accuracies or training sample size between the balanced and unbalanced studies in the held-out test group.

Because the ADHD-200 dataset was the main data source, most studies (N=26) used resting-state fMRI data (rs-fMRI), or rs-fMRI in combination with sMRI data (multi-modal, N=14). Only seven studies used sMRI data, and only and two used task-based fMRI data. The sample sizes for the multi-modal and rs-fMRI studies were significantly larger than those of the sMRI and task-based fMRI studies ($F_{(3, 47)}$= 4.3, p=0.009, Supplementary Figure 4A). However, except for the two task-based fMRI studies, which both used LOOCV and reported significantly lower accuracies than other MRI modalities (t=-23.3, p<.0001), there was no difference in reported classification accuracies observed among the sMRI, rs-fMRI, or multi-modal studies (Supplementary Figure 4B).

The ADHD-200 dataset has a mix of children, adolescents and young adults (age 7-21). Ten other studies focused only on children and/or adolescents (under age 18). Only three studies examined classification models for older adults (Chaim-Avancini et al., 2017; Wang et al., 2013; Yao et al., 2018). Overall, the difference in accuracy across the three types of age compositions was not significant ($F_{(2, 47)}$ = 2.64, p= 0.08).

Most studies used a mixture of male and female participants. Four studies only included boys (Johnston et al., 2014; Lim et al., 2013; Yao et al., 2018; Zhu et al., 2008). These four reported significantly higher classification accuracies than all other studies that used a mixture of males and females ($F_{(1, 47)}$ =10.06, p = 0.003). However, all four were small studies (n=20~189). Three reported LOOCV and one reported 10-Fold-CV accuracies.

Across all studies, the most frequently used model was the support vector machine (SVM). It was used in 20 (41%) studies. SVM, and most other ML models cannot directly analyze images. Instead, they analyze some transformation of images such as regional volumes or cortical thickness. In contrast, convolutional neural networks (CNNs) can analyze images directly and thus have access to all the information available. Only in recent years (2017-2020) have studies applied CNN methods to MRI images (N=6). We did not find any statistically significant differences between the accuracies reported with SVM, CNN, or other models for ADHD $F_{(2, 48)}$ = 0.77, p = 0.47, Supplementary Figure 5).

Discussion and Qualitative Review

Our quantitative analysis of prediction accuracy for ADHD revealed several significant findings. First, accuracies based on K-Fold-CV or LOOCV were significantly higher than those reported using held-out tests, which suggests that CV methods may over-estimate model performance. Second, we found greater variability of test accuracies reported in studies with small sample sizes than those of larger sample sizes and an inverse relationship of sample size and K-Fold-CV accuracies. Third, estimates of accuracy increased with publication year. This effect was driven primarily by the held-out test accuracies. Since sample size has been roughly the same since 2012, we believe that the increasing accuracy over time was due to several design features: 1) the use of more sophisticated models (such as deep neural networks and CNN models), 2) improved methods of data balancing and data augmentation, and 3) use of feature selection and feature space reduction methods. We found no significant effects of MRI feature modality or type of ML model. We discuss the implications of these findings and provide further review of some study characteristics that were not examined in our quantitative analysis.

Cross-Validation vs Held-out Test Set

In the CV approach, the validation samples used to estimate accuracy are not used during the model training/fitting at each iteration. They are, nevertheless, used as training examples in other iterations. Moreover, because there are many iterations, the validation set can influence parameter estimates. In contrast, the held-out test method uses a test set that was never used during model training. As a result, CV accuracies have been shown to overestimate test set accuracy when both are available (Brown et al., 2012; Dai et al., 2012). Our results confirm the inflation of accuracy by K-Fold-CV or LOOCV. Held-out test accuracy is a better indicator of model performance with unseen samples.

More than half the studies (N=27, 55%) reported only CV results without a held-out test set. An earlier review reported 13% of ADHD neuroimaging (including MRI and electroencephalographic) studies consisted of “circular analysis”, where independent test sets were not used (Pulini, Kerr, Loo, & Lenartowicz, 2019). Our results are more similar to what Kriegeskorte et al. (2009) had estimated, 42%~56% of studies consisted of “circular analysis”, based on all fMRI studies published in five prestigious journals (Nature, Science, Nature Neuroscience, Neuron, Journal of Neuroscience) in 2008. Nevertheless, our review highlights the importance of building a large dataset through collaborations and open data sharing as we pointed out that the majority of the studies that were able to afford a held-out test were those that used the ADHD-200 dataset.

Sample Size

Machine learning, particularly deep learning, often requires large sample sizes due to the large number of parameters and hyperparameters that a model needs to learn. However, many neuroimaging studies of ADHD used very small sample sizes. In our small sample group, the sample size ranged from 20 to 239 (average sample size 112). Small sample sizes can lead to model overfitting and overestimates of accuracy (Brain & Webb, 1999; Wolfers, Buitelaar, Beckmann, Franke, & Marquand, 2015) . In our review, this effect was reflected in the large variability of accuracies in the CV studies. Indeed, some of the highest and lowest test accuracies were reported in studies with extremely small sample sizes. None of the studies reviewed here used a learning curve analysis to assess overfitting. This method, which examines the relationship of model performance over various numbers of training sample sizes, can help us to determine if a model is overfit and if it can benefit from more training examples (Zhang-James et al., 2020).

We found a negative relationship between sample size and K-Fold-CV accuracies. Because increasing the number of training samples typically improves performance (Bengio, Courville, & Vincent, 2013), this suggests that the lower estimates of accuracy from the larger samples are more likely to be correct than the higher estimates from smaller samples. Those higher estimates were likely biased as described in the prior section. Pulini et al (2019) also reported a negative relationship of sample size and accuracies in ADHD imaging studies and Vabalas observed the negative relationship of sample size on reported accuracies in machine learning classifiers of autism spectrum disorders (Vabalas, Gowen, Poliakoff, & Casson, 2019). Both reviews were based on studies with sample sizes up to only ~1,000. Similar observations were also made by Wolfers and colleagues when reviewing neuroimaging-based diagnostics for a number of different psychiatric disorders (Wolfers et al., 2015).

Sample Heterogeneity and Data Imbalance.

Although collaborative consortia, such as the ADHD-200, used relatively large samples sizes, such collaborations raise issues about sample heterogeneity and the use of imbalanced data. For example, like many other clinically referred samples, the ADHD-200 dataset had more boys than girls in the ADHD group compared with the control group. The ADHD group also had lower IQs than the control group. In addition, the demographic composition and sample acquisition methods differed across different study sites. The problem of dataset imbalance was addressed by several participating teams. Brown and colleagues from the University of Alberta found that models using only demographic information including age, sex, handedness, and IQ had sufficient statistical power to achieve a test accuracy 62.5%, higher than their models using fMRI features (Brown et al., 2012). In the work of Colby et al. (2012), a model using only demographic information had a higher accuracy (62.7%) than models using multimodal MRI features (55%). Both models using only the demographic features, although not meeting the requirements of the competition, outperformed the winning team that reported 61% accuracy using both structural and rs-fMRI data along with the demographic predictors in an ensemble model (Eloyan et al., 2012). An additional study by Sidhu and colleagues also reported better accuracy using demographic information than the rs-fMRI features using the ADHD-200 dataset (Sidhu, Asgarian, Greiner, & Brown, 2012). These observations highlight the concerns of data imbalance, and suggest that, if not dealt with carefully, the classifiers could be learning the neural correlates of the demographic features, rather than the diagnostic groups.

Some studies used methods to address the problem of unbalanced data. One approach is random undersampling, i.e, removing some research participants and creating a smaller sample size that is balanced for confounding factors (M. N. I. Qureshi et al., 2016; Wang et al., 2013). This is in contrast to oversampling, where some random samples from the minority classes were duplicated to create a lager and balanced dataset. Others used regression to control confounding factors such as age, sex, and acquisition sites, and used adjusted MRI features (residuals) in the classification algorithms (Deshpande et al., 2015; Fair et al., 2012). Some studies mentioned data balancing, but did not provide details on how it was done (Ghiassian et al., 2016). Lim et al (2013) used a gaussian process classifier to discriminate 29 boys with ADHD from 19 control boys. The limited samples sizes prohibited subsampling to balance the data. They noted, although the boys with ADHD had significantly lower IQ than the control boys, the model-generated probability of having ADHD was not correlated with IQ, age, and other clinical features (Lim et al., 2013). In more recent studies, more sophisticated methods such as Synthetic Minority Over-sampling Technique (SMOTE, (Chawla, Bowyer, Hall, & Kegelmeyer, 2002)) were used to generate synthetic minority samples to combat the sample imbalance problem (Riaz, Asad, Alonso, et al., 2018).

Previous studies from other fields have shown that not all the class balancing methods work equally well in reducing classifiers' bias towards the majority class and guarantee good performance (Blagus & Lusa, 2013; He & Ma, 2013). In the ADHD studies that we reviewed, we did not find found higher held-out test accuracies in balanced studies than the unbalanced studies. Balanced studies in the K-Fold-CV group reported higher accuracies than those with unbalanced samples. However, they were all smaller studies than the unbalanced studies. It is not clear, at least for the K-Fold-CV group, if balanced designs led to higher accuracies, because sample size was a strong and negative predictor for the accuracies. Nevertheless, higher accuracies in the balanced LOOCV group, independent of the sample size, do suggest that sample balancing may have helped model performance to some degree. More studies and larger sample sizes will be needed to find the appropriate class balancing methods and assess the potential benefit.

Classification performance metrics

When test sets (or cross-validation sets) are also imbalanced, the overall accuracy may not be an ideal indicator of the performance of the classifier. A high accuracy can simply result from a classifier that classifies all samples into the class that has more participants. Most studies (N= 36, 75%) addressed this concern by also reporting sensitivity (True Positive rate, TP, the percentage of correctly identified cases (ADHD)) and specificity (True Negative rate, TN, the percentage of correctly identified controls). Three studies reported balanced accuracy, which is the arithmetic mean of the sensitivity and specificity; and three studies reported Youden's J-statistic (sensitivity + specificity -1).

Compared with percent correct, a better method to evaluate the overall performance of a classifier is the area under the Receiver Operating Characteristic curve (ROC (Fawcett, 2004)). The ROC curve plots sensitivity over the full range of false positive rates (equivalent to 1-specificity). The area under the ROC (AUC) measures the overall diagnostic accuracy of a classifier. Higher AUCs indicate better discriminating power (with 1 for the perfect classifier and 0.5 for the random non-discriminative classifier). The AUC is in general less sensitive to imbalance of a dataset compared with the percent correct measure, because AUC does not have bias toward models that perform well on the majority class at the expense of the minority class (He & Ma, 2013). Davis et al (2006) suggested that the area under the Precision-recall curve (AUPRC) is superior for assessing extremely imbalanced datasets and more informative than the ROC curve. The AUPRC plots precision (the percentage of examples classified as positive that are true positive, also known as positive predictive value, PPV) over recall (sensitivity). Overall, in the body of literature that we examined, no studies reported the AUPRC, and only 13 reported the AUC.

Other popular metrics for machine learning models are F1-score and Matthew's correlation coefficient (MCC). The F1-score is the harmonic mean of precision and recall. MCC is the correlation coefficient between the predicted and actual classes. Like the areas under the PRC or ROC, both the F1-score and MCC are better indicators of model performance than the percent correct statistic if test data classes are imbalanced. However, of the studies included in this review, only three reported F1-scores, and only two reported MCC.

Because most studies used percent correct to measures accuracy, we could only analyze percent correct in the current review. This may not represent the true model performance due to the limitations of this metric. We recommend that future studies adopt ROC or PRC analysis methods. Furthermore, inspecting the curves visually can reveal more information about how well the models discriminate classes at different decision thresholds. We don't recommend metrics such as the F1-score, MCC, and J-statistics, because these scores only capture the diagnostic matrix at a single threshold level. Furthermore, the performance metrics are important, not only for properly interpreting test results, but also for model training. If a model was trained by maximizing a biased metric, it will not be fully optimized to generalize to other samples. Finally, metrics that are insensitive to *class* imbalance (such as AUC or AUPRC) do not protect against biased due to feature imbalance as discussed in the prior section.

Age and Sex

Although ADHD onsets prior to age 12, two-thirds of children continue to have symptoms and functional impairments into adulthood (Faraone, Biederman, & Mick, 2006). Longitudinal data show that some ADHD-associated brain alterations diminish during adolescence and adulthood (Castellanos et al., 2002; Shaw et al., 2014; Shaw et al., 2006). Consistent with this, the very large ENIGMA ADHD study reported significant ADHD vs. control differences for children but not for adolescents and adults (Hoogman et al., 2017; Hoogman et al., 2019). Neuroimaging classifiers studies have focused on younger populations; only three studies developed ML classifiers for adults. Our observation of a lack of difference in predictive accuracy between classifiers for children/adolescents vs. adults is, therefore, inconclusive due to the small numbers of the adult studies. Few longitudinal studies have been reported for imaging in ADHD (Castellanos et al., 2002; Shaw et al., 2014; Shaw et al., 2006). No machine learning models have been applied to longitudinal data yet. More efforts are needed to overcome the shortage of adult ADHD samples, as well as imaging data across the life span.

ADHD is more prevalent in boys than in girls (Faraone et al., 2015; Ottosen et al., 2019). As a result, although the majority of the studies included samples from the both males and females, a high percentage of ADHD samples were from males (i.e., ~80% male in ADHD-200 dataset). However, the control samples were often balanced (i.e., 52% male in ADHD-200 dataset). If sex is left unbalanced, it could result in erroneous prediction results, as we described in the above sections. Furthermore, brain alterations have been found to differ between the sexes at different ages (Almeida Montes et al., 2013; Hoogman et al., 2019; Onnink et al., 2014). The low representation of females in available samples may prevent the classifiers from learning female-specific brain alternations. Our quantitative analysis showed significantly higher accuracies in four male-only studies than other studies of sex-mixed samples. However, all four were studies with small sample sizes (<189), with three reporting LOOCV accuracies and the other reporting 10-Fold CV. Given the sample size effect and inflation by CV methods, it is inconclusive if ML models predict ADHD better in boys than girls.

MRI Modality

Although we found no significant difference in the accuracies reported for the sMRI and rs-fMRI studies, the small number of studies using sMRI data preclude any meaningful inferences regarding which MRI modality is the most informative for discriminating ADHD patients from controls. Some studies attempted to identify the most informative MRI data modality. Qureshi et al (2017) found that sMRI features yielded the highest prediction accuracy. Colby et al. (2012) found that combined multi-modality features performed best compared with individual data modalities. However, all the MRI models performed worse than a classifier using only demographic features (Colby et al., 2012). In a later study using a three-dimensional CNN model, Zou and colleagues extracted higher-level features from the sMRI and rs-fMRI modalities separately. This design leveraged the relationship between the two MRI modalities, yet still was able to extract independent features that collectively were useful for classification (Zou, Zheng, Miao, Mckeown, & Wang, 2017). The authors also found that using multi-modal features outperformed either data modality alone (Zou et al., 2017). Despite these individual observations, the overall lack of statistically significant differences in accuracies across different modalities in our review suggests that more studies are needed before we can determine which MRI modalities or combinations thereof are most informative for diagnostic classification.

ML Classifiers.

SVM was the ML model that was used most frequently, accounting for 41% of studies. SVM, however, is limited in handling images and relies on other preprocessing methods to extract a tabular representation of three-dimensional brain images. In more recent years, an increasing number of studies have used CNNs, which were developed for image analysis. We did not, however, observe statistically significant differences between the accuracies of the SVM and CNN models for ADHD. This finding is limited by the small number of studies using CNN classifiers. Nevertheless, because the use of CNNs will likely increase in the future, we here describe their current contributions to the field and their potential for the future.

Riaz et al. (2017) used a CNN-based method (FCNet) to extract the functional connectivity (FC) of brain regions and then trained an SVM classifier using the extracted features to discriminate ADHD from control participants. The classifier achieved a highest held-out test accuracy of 68.6% for the ADHD-200 Peking subset. In the follow up study, the team built an end-to-end model system, DeepFMRI, which utilized multiple FCNets to extract features that were then fed into a deep neural network (Riaz, Asad, Arif, et al., 2018). DeepFMRI streamlined the feature generation and selection as well as classification in one framework, and achieved a highest test accuracy of 73.1% for the NYU subset. Using preprocessed rs-fMRI and sMRI features as independent inputs, Zou et al. used a two-branched three-dimensional CNN to learn hierarchical features from each unique modality in a joint learning task. The multi-modal joint learning CNN architecture was superior to CNNs using either data modality alone (Zou et al., 2017). Aradhya et al. (2019) also used a CNN classifier and extracted features using the Deep Transformation Method (DTM).

Most studies, including many CNN studies, used pre-processed MRI features, such as those anatomized to an AAL template. Mao and colleagues argued that rather than using hand-crafted features, one should use a CNN to directly learn discriminatory features from images. Their four-dimensional CNN classifier, designed to learn and extract spatial and temporal features from rs-fMRI images, discriminated ADHD from control participants with an accuracy of 71.3% (Mao et al., 2019). To increase their sample size and reduce overfitting, the authors augmented data by transforming rs-fMRI data into many short and fixed-lengthed video clips. Despite their promising results, they acknowledged that much work is still needed to localize the most discriminative sequences. Interestingly, a CNN using activation correlations from individual brain regions of the Default Mode Network (DMN) of the brain outperformed those using whole brain features (Ariyarathne, Silva, Dayarathna, Meedeniya, & Jayarathne, 2020). Using only one relevant brain region substantially reduced feature space and complexity. The significantly improved model performance also suggests that current sample sizes, in relation to the number of features available, maybe limiting the CNN models' capacity. With more samples becoming available in the future, and the increased datasets of publicly available raw MRI images, CNN methods will likely to be seen in more and more studies and be explored to their full capacity for feature extraction and classification as has been the case for computer vision (Arcadu et al., 2019; Bhanumathi & Sangeetha, 2019; Iqbal, Ghani, Saba, & Rehman, 2018; Lin et al., 2018; Toyonaga et al., 2017).

Building Larger Datasets

Sample size has been a major bottleneck impeding the development of more accurate and clinically useful imaging classifiers for ADHD. The largest MRI dataset, to date, has been built by the Enhancing Neuro Imaging Genetics Through Meta-Analysis (ENIGMA) consortium. Under the umbrella of the ENIGMA consortium, many independent working groups for specific diseases or phenotypes have been established, including ADHD. By implementing standardized data processing protocols and pipelines, the ENIGMA consortium made it possible to share data across many sites to perform within-disorder and cross-disorder studies (Boedhoe et al., 2020; Paul M. Thompson et al., 2017; P. M. Thompson et al., 2020; P. M. Thompson et al., 2014) . The ENIGMA ADHD Working Group has obtained over 4,100 samples of ADHD participants and controls from 37 sites thus far. In the initial ENIGMA ADHD reports, Hoogman and colleagues reported that, for children, ADHD was associated with significant volumetric reductions in intracranial volume, amygdala, caudate nucleus, nucleus accumbens, hippocampus, and cortical surface areas from many regions (Hoogman et al., 2017; Hoogman et al., 2019). No significant differences were found for adolescents or adults. Furthermore, the estimated effect sizes for children were small, ranging from 0.11 to 0.19. Users of the ENIGMA ADHD dataset, however, face the same problems of data heterogeneity and imbalanced demographic groups as those using the ADHD-200 dataset. Significant challenges remain when using such data to build a machine learning classifier. Furthermore, the ENIGMA ADHD data is primarily preprocessed sMRI data in tabular form. Not all sites have data on other modalities, such as rs-fMRI or DTI, available for their samples. The ENIGMA ADHD sites have not yet pooled raw MRI images, which is needed for CNN models.

Conclusions

Our review of ML studies of MRI-based ADHD diagnostic classifiers has important implications for methods development, but these studies have not yet led to clinically useful classifiers. Our review shows that the variability of results across studies is due, in part, to differences in methodology. Future work should use the largest samples possible and should rely on a held-out test set, rather than cross-validation for estimating prediction accuracy. Future studies should not rely on percent correct as a measure of accuracy in unbalanced samples. Our analysis also highlighted the need of data from underrepresented groups, particularly females and adults. We hope that our review provides a better understanding of the efforts invested in developing ADHD imaging classifiers in the field and encourages more stringent model design and data processing for future studies. In the meanwhile, the initial results from the ENIGMA ADHD consortium should encourage more sites to participate. The lack of a very large multi-modal dataset that include sufficient data from both sex and all ages may be the biggest impediment to developing a clinically useful classifier for diagnosing ADHD.

Funding Acknowledgement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. However, we acknowledge the following funding agencies who support the authors.

Dr. Zhang-James is supported by the European Union’s Seventh Framework Programme for research, technological development and demonstration (grant number 602805) and the European Union’s Horizon 2020 research and innovation programme (grant number 667302).

Dr. Hoogman is supported by a personal Veni grant from of the Netherlands Organization for Scientific Research (NWO, grant number 91619115)

Dr. Franke is supported by a personal Vici grant (grant number 016-130-669) from the Netherlands Organization for Scientific Research (NWO).

Dr. Faraone is supported by the European Union's Seventh Framework Programme for research, technological development and demonstration (grant number 602805), the European Union's Horizon 2020 research and innovation programme (grant number 667302 & 728018) and United States National Institute of Mental Health (NIMH, grant numbers 5R01MH101519 and U01 MH109536-01).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723. doi:10.1109/TAC.1974.1100705

Almeida Montes, L. G., Prado Alcantara, H., Martinez Garcia, R. B., De La Torre, L. B., Avila Acosta, D., & Duarte, M. G. (2013). Brain cortical thickness in ADHD: age, sex, and clinical correlations. *J Atten Disord, 17*(8), 641-654. doi:10.1177/1087054711434351

1087054711434351 [pii]

Aradhya, A., MS., Joglekar, A., S, S., & Pratama, M. (2019). Deep Transformation Method for Discriminant Analysis of Multi-Channel Resting State fMRI *Proceedings of the AAAI Conference on Artificial Intelligence, 33*, 01.

Arcadu, F., Benmansour, F., Maunz, A., Willis, J., Haskova, Z., & Prunotto, M. (2019). Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *npj Digital Medicine, 2*(1), 92. doi:10.1038/s41746-019-0172-3

Ariyarathne, G., Silva, S. D., Dayarathna, S., Meedeniya, D., & Jayarathne, S. (2020). *ADHD Identification using Convolutional Neural Network with Seed-based Approach for fMRI Data*. Paper presented at the Proceedings of the 2020 9th International Conference on Software and Computer Applications, Langkawi, Malaysia. https://doi.org/10.1145/3384544.3384552

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell, 35*(8), 1798-1828. doi:10.1109/TPAMI.2013.50

Bhanumathi, V., & Sangeetha, R. (2019, 15-16 March 2019). *CNN Based Training and Classification of MRI Brain Images.* Paper presented at the 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS).

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics, 14*(1), 106. doi:10.1186/1471-2105-14-106

Boedhoe, P. S. W., van Rooij, D., Hoogman, M., Twisk, J. W. R., Schmaal, L., Abe, Y., . . . van den Heuvel, O. A. (2020). Subcortical Brain Volume, Regional Cortical Thickness, and Cortical Surface Area Across Disorders: Findings From the ENIGMA ADHD, ASD, and OCD Working Groups. *Am J Psychiatry, 177*(9), 834-843. doi:10.1176/appi.ajp.2020.19030331

Brain, D., & Webb, G. I. (1999). *On The Effect of Data Set Size on Bias And Variance in Classification Learning*. Paper presented at the Proceedings of the Fourth Australian Knowledge Acquisition Workshop (AKAW '99), Sydney.

Brown, M. R., Sidhu, G. S., Greiner, R., Asgarian, N., Bastani, M., Silverstone, P. H., . . . Dursun, S. M. (2012). ADHD-200 Global Competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. *Front Syst Neurosci, 6*, 69. doi:10.3389/fnsys.2012.00069

Bruchmuller, K., Margraf, J., & Schneider, S. (2012). Is ADHD diagnosed in accord with diagnostic criteria? Overdiagnosis and influence of client gender on diagnosis. *J Consult Clin Psychol, 80*(1), 128-138. doi:10.1037/a0026582

2011-30100-001 [pii]

Castellanos, F. X., Lee, P. P., Sharp, W., Jeffries, N. O., Greenstein, D. K., Clasen, L. S., . . . Rapoport, J. L. (2002). Developmental trajectories of brain volume abnormalities in children and adolescents with attention-deficit/hyperactivity disorder. *JAMA, 288*(14), 1740-1748. doi:10.1001/jama.288.14.1740

Chaim-Avancini, T. M., Doshi, J., Zanetti, M. V., Erus, G., Silva, M. A., Duran, F. L. S., . . . Busatto, G. F. (2017). Neurobiological support to the diagnosis of ADHD in stimulant-naïve adults: pattern recognition analyses of MRI data. *Acta Psychiatrica Scandinavica, 136*(6), 623-636. doi:10.1111/acps.12824

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique*. Paper presented at the Journal Of Artificial Intelligence Research.

Colby, J. B., Rudie, J. D., Brown, J. A., Douglas, P. K., Cohen, M. S., & Shehzad, Z. (2012). Insights into multimodal imaging classification of ADHD. *Front Syst Neurosci, 6*, 59. doi:10.3389/fnsys.2012.00059

Consortium, A.-. (2012). The ADHD-200 Consortium: A Model to Advance the Translational Potential of Neuroimaging in Clinical Neuroscience. *Frontiers in systems neuroscience, 6*, 62. doi:10.3389/fnsys.2012.00062

Corkum, P. V., & Siegel, L. S. (1993). Is the Continuous Performance Task a valuable research tool for use with children with Attention-Deficit-Hyperactivity Disorder? *J Child Psychol Psychiatry, 34*(7), 1217-1239. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8245143

Dai, D., Wang, J., Hua, J., & He, H. (2012). Classification of ADHD children through multimodal magnetic resonance imaging. *Front Syst Neurosci, 6*, 63. doi:10.3389/fnsys.2012.00063

Dalsgaard, S., Ostergaard, S. D., Leckman, J. F., Mortensen, P. B., & Pedersen, M. G. (2015). Mortality in children, adolescents, and adults with attention deficit hyperactivity disorder: a nationwide cohort study. *Lancet, 385*(9983), 2190-2196. doi:10.1016/S0140-6736(14)61684-6

Dane, A. V., Schachar, R. J., & Tannock, R. (2000). Does actigraphy differentiate ADHD subtypes in a clinical research setting? *Journal of the American Academy of Child and Adolescent Psychiatry, 39*(6), 752-760. Retrieved from http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=0010846310

Davis, J., & Goadrich, M. (2006). *The relationship between Precision-Recall and ROC curves*. Paper presented at the Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, USA. https://doi.org/10.1145/1143844.1143874

Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., . . . Neale, B. M. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat Genet, 51*(1), 63-75. doi:10.1038/s41588-018-0269-7

Deshpande, G., Wang, P., Rangaprakash, D., & Wilamowski, B. (2015). Fully Connected Cascade Artificial Neural Network Architecture for Attention Deficit Hyperactivity Disorder Classification From Functional Magnetic Resonance Imaging Data. *IEEE Trans Cybern, 45*(12), 2668-2679. doi:10.1109/TCYB.2014.2379621

Dickstein, S. G., Bannon, K., Castellanos, F. X., & Milham, M. P. (2006). The neural correlates of attention deficit hyperactivity disorder: an ALE meta-analysis. *J Child Psychol Psychiatry, 47*(10), 1051-1062. doi:10.1111/j.1469-7610.2006.01671.x

Du, J., Wang, L., Jie, B., & Zhang, D. (2016). Network-based classification of ADHD patients using discriminative subnetwork selection and graph kernel PCA. *Comput Med Imaging Graph, 52*, 82-88. doi:10.1016/j.compmedimag.2016.04.004

Eloyan, A., Muschelli, J., Nebel, M. B., Liu, H., Han, F., Zhao, T., . . . Caffo, B. (2012). Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Front Syst Neurosci, 6*, 61. doi:10.3389/fnsys.2012.00061

Fair, D. A., Nigg, J. T., Iyer, S., Bathula, D., Mills, K. L., Dosenbach, N. U., . . . Milham, M. P. (2012). Distinct neural signatures detected for ADHD subtypes after controlling for micro-movements in resting state functional connectivity MRI data. *Front Syst Neurosci, 6*(80), 80. doi:10.3389/fnsys.2012.00080

Faraone, S. V. (2005). The scientific foundation for understanding attention-deficit/hyperactivity disorder as a valid psychiatric disorder. *Eur Child Adolesc Psychiatry, 14*, 1-10.

Faraone, S. V., Asherson, P., Banaschewski, T., Biederman, J., Buitelaar, J. K., Ramos-Quiroga, J. A., . . . Franke, B. (2015). Attention-deficit/hyperactivity disorder. *Nat Rev Dis Primers, 1*, 15020. doi:10.1038/nrdp.2015.20

Faraone, S. V., & Biederman, J. (2000). Nature, nurture, and attention deficit hyperactivity disorder. *Developmental Review, 20*, 568-581.

Faraone, S. V., Biederman, J., & Mick, E. (2006). The age-dependent decline of attention deficit hyperactivity disorder: a meta-analysis of follow-up studies. *Psychol Med, 36*(2), 159-165. doi:10.1017/S003329170500471X

Faraone, S. V., Bonvicini, C., & Scassellati, C. (2014). Biomarkers in the diagnosis of ADHD--promising directions. *Curr Psychiatry Rep, 16*(11), 497. doi:10.1007/s11920-014-0497-1

Faraone, S. V., Newcorn, J. H., Antshel, K. M., Adler, L., Roots, K., & Heller, M. (2016). The Groundskeeper Gaming Platform as a Diagnostic Tool for Attention-Deficit/Hyperactivity Disorder: Sensitivity, Specificity, and Relation to Other Measures. *J Child Adolesc Psychopharmacol, 26*(8), 672-685. doi:10.1089/cap.2015.0174

Fawcett, T. (2004). ROC Graphs: Notes and Practical Considerations for Researchers. *Machine Learning, 31*, 1-38.

Franke, B., Michelini, G., Asherson, P., Banaschewski, T., Bilbow, A., Buitelaar, J. K., . . . Reif, A. (2018). Live fast, die young? A review on the developmental trajectories of ADHD across the lifespan. *Eur Neuropsychopharmacol, 28*(10), 1059-1088. doi:10.1016/j.euroneuro.2018.08.001

Ghiassian, S., Greiner, R., Jin, P., & Brown, M. R. G. (2016). Using functional or structural magnetic resonance images and personal characteristic data to identify ADHD and autism. *PLoS One, 11*(12). doi:10.1371/journal.pone.0166934

Ginsberg, Y., Quintero, J., Anand, E., Casillas, M., & Upadhyaya, H. P. (2014). Underdiagnosis of attention-deficit/hyperactivity disorder in adult patients: a review of the literature. *Prim Care Companion CNS Disord, 16*(3). doi:10.4088/PCC.13r01600

Hamshere, M. L., Langley, K., Martin, J., Agha, S. S., Stergiakouli, E., Anney, R. J., . . . Thapar, A. (2013). High loading of polygenic risk for ADHD in children with comorbid aggression. *Am J Psychiatry, 170*(8), 909-916. doi:1680039 [pii]

10.1176/appi.ajp.2013.12081129

Hart, H., Chantiluke, K., Cubillo, A. I., Smith, A. B., Simmons, A., Brammer, M. J., . . . Rubia, K. (2014). Pattern classification of response inhibition in ADHD: toward the development of neurobiological markers for ADHD. *Hum Brain Mapp, 35*(7), 3083-3094. doi:10.1002/hbm.22386

He, H., & Ma, Y. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications*: Wiley-IEEE Press.

Homack, S., & Riccio, C. A. (2006). Conners' Continuous Performance Test (2nd ed.; CCPT-II). *J Atten Disord, 9*(3), 556-558. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16481673

Hoogman, M., Bralten, J., Hibar, D. P., Mennes, M., Zwiers, M. P., Schweren, L. S. J., . . . Franke, B. (2017). Subcortical brain volume differences in participants with attention deficit hyperactivity disorder in children and adults: a cross-sectional mega-analysis. *Lancet Psychiatry, 4*(4), 310-319. doi:10.1016/S2215-0366(17)30049-4

Hoogman, M., Muetzel, R., Guimaraes, J. P., Shumskaya, E., Mennes, M., Zwiers, M. P., . . . Franke, B. (2019). Brain Imaging of the Cortex in ADHD: A Coordinated Analysis of Large-Scale Clinical and Population-Based Samples. *Am J Psychiatry, 176*(7), 531-542. doi:10.1176/appi.ajp.2019.18091033

Iannaccone, R., Hauser, T. U., Ball, J., Brandeis, D., Walitza, S., & Brem, S. (2015). Classifying adolescent attention-deficit/hyperactivity disorder (ADHD) based on functional and structural imaging. *Eur Child Adolesc Psychiatry, 24*(10), 1279-1289. doi:10.1007/s00787-015-0678-4

Iqbal, S., Ghani, M. U., Saba, T., & Rehman, A. (2018). Brain tumor segmentation in multi-spectral MRI using convolutional neural networks (CNN). *Microsc Res Tech, 81*(4), 419-427. doi:10.1002/jemt.22994

Johnston, B. A., Mwangi, B., Matthews, K., Coghill, D., Konrad, K., & Steele, J. D. (2014). Brainstem abnormalities in attention deficit hyperactivity disorder support high accuracy individual diagnostic classification. *Hum Brain Mapp, 35*(10), 5179-5189. doi:10.1002/hbm.22542

Joseph, N., Zhang-James, Y., Perl, A., & Faraone, S. V. (2015). Oxidative Stress and Attention Deficit Hyperactivity Disorder: A Meta-Analysis. *J Atten Disord, 19*(11), 915-924. doi:10.1177/1087054713510354

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci, 12*(5), 535-540. doi:10.1038/nn.2303

Lambert, N. M., & Hartsough, C. S. (1998). Prospective study of tobacco smoking and substance dependencies among samples of ADHD and non-ADHD participants. *J Learn Disabil, 31*(6), 533-544.

Levene, H. (1960). Robust tests for equality of variances. In S. G. G. I. Olkin, W. Hoeffding, W. G. Madow, and H. B. Mann (Ed.), *In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 278-292). Menlo Park, CA: Stanford University Press.

Lichtenstein, P., Halldner, L., Zetterqvist, J., Sjolander, A., Serlachius, E., Fazel, S., . . . Larsson, H. (2012). Medication for attention deficit-hyperactivity disorder and criminality. *N Engl J Med, 367*(21), 2006-2014. doi:10.1056/NEJMoa1203241

Lim, L., Marquand, A., Cubillo, A. A., Smith, A. B., Chantiluke, K., Simmons, A., . . . Rubia, K. (2013). Disorder-specific predictive classification of adolescents with attention deficit hyperactivity disorder (ADHD) relative to autism using structural magnetic resonance imaging. *PLoS One, 8*(5), e63660. doi:10.1371/journal.pone.0063660

Lin, W., Tong, T., Gao, Q., Guo, D., Du, X., Yang, Y., . . . Alzheimer's Disease Neuroimaging, I. (2018). Convolutional Neural Networks-Based MRI Image Analysis for the Alzheimer's Disease Prediction From Mild Cognitive Impairment. *Front Neurosci, 12*, 777. doi:10.3389/fnins.2018.00777

Mackie, S., Shaw, P., Lenroot, R., Pierson, R., Greenstein, D. K., Nugent, T. F., 3rd, . . . Rapoport, J. L. (2007). Cerebellar development and clinical outcome in attention deficit hyperactivity disorder. *Am J Psychiatry, 164*(4), 647-655. doi:10.1176/ajp.2007.164.4.647

Mao, Z., Su, Y., Xu, G., Wang, X., Huang, Y., Yue, W., . . . Xiong, N. (2019). Spatio-temporal deep learning method for ADHD fMRI classification. *Information Sciences, 499*, 1-11. doi:10.1016/j.ins.2019.05.043

Onnink, A. M., Zwiers, M. P., Hoogman, M., Mostert, J. C., Kan, C. C., Buitelaar, J., & Franke, B. (2014). Brain alterations in adult ADHD: effects of gender, treatment and comorbid depression. *Eur Neuropsychopharmacol, 24*(3), 397-409. doi:10.1016/j.euroneuro.2013.11.011

S0924-977X(13)00342-8 [pii]

Ottosen, C., Larsen, J. T., Faraone, S. V., Chen, Q., Hartman, C., Larsson, H., . . . Dalsgaard, S. (2019). Sex Differences in Comorbidity Patterns of Attention-Deficit/Hyperactivity Disorder. *J Am Acad Child Adolesc Psychiatry, 58*(4), 412-422.e413. doi:10.1016/j.jaac.2018.07.910

Peng, X., Lin, P., Zhang, T., & Wang, J. (2013). Extreme learning machine-based classification of ADHD using brain structural MRI data. *PLoS One, 8*(11), e79476. doi:10.1371/journal.pone.0079476

Pulini, A. A., Kerr, W. T., Loo, S. K., & Lenartowicz, A. (2019). Classification Accuracy of Neuroimaging Biomarkers in Attention-Deficit/Hyperactivity Disorder: Effects of Sample Size and Circular Analysis. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 4*(2), 108-120. doi:10.1016/j.bpsc.2018.06.003

Qureshi, M. N., Min, B., Jo, H. J., & Lee, B. (2016). Multiclass classification for the differential diagnosis on the ADHD subtypes using recursive feature elimination and hierarchical extreme learning machine: Structural MRI study. *PLoS One, 11*(8), e0160697. doi:10.1371/journal.pone.0160697

PONE-D-16-09099 [pii]

Qureshi, M. N. I., Min, B., Jo, H. J., & Lee, B. (2016). Multiclass classification for the differential diagnosis on the ADHD subtypes using recursive feature elimination and hierarchical extreme learning machine: Structural MRI study. *PLoS One, 11*(8). doi:10.1371/journal.pone.0160697

Qureshi, M. N. I., Oh, J., Min, B., Jo, H. J., & Lee, B. (2017). Multi-modal, multi-measure, and multi-class discrimination of ADHD with hierarchical feature extraction and extreme learning machine using structural and functional brain MRI. *Front Hum Neurosci, 11*. doi:10.3389/fnhum.2017.00157

Reiersen, A., & Todorov, A. (2013). Exploration of ADHD Subtype Definitions and Co-Occurring Psychopathology

in a Missouri Population-Based Large Sibship Sample. *Scandinavian Journal of Child and Adolescent Psychiatry and Psychology, 1*(1), 3-13

.

Riaz, A., Asad, M., Alonso, E., & Slabaugh, G. (2018). Fusion of fMRI and non-imaging data for ADHD classification. *Comput Med Imaging Graph, 65*, 115-128. doi:10.1016/j.compmedimag.2017.10.002

Riaz, A., Asad, M., Arif, S. M. M. R. A., Alonso, E., Dima, D., Corr, P., & Slabaugh, G. (2017). *FCNet: A Convolutional Neural Network for Calculating Functional Connectivity from Functional MRI.* Paper presented at the Connectomics in NeuroImaging. CNI 2017. Lecture Notes in Computer Science. .

Riaz, A., Asad, M., Arif, S. M. M. R. A., Alonso, E., Dima, D., Corr, P., & Slabaugh, G. (2018). *Deep fMRI: AN end-to-end deep network for classification of fMRI data*. Paper presented at the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC.

Riccio, C. A., & Reynolds, C. R. (2001). Continuous performance tests are sensitive to ADHD in adults but lack specificity. A review and critique for differential diagnosis. *Ann N Y Acad Sci, 931*, 113-139. Retrieved from http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=11462737

Riglin, L., Collishaw, S., Thapar, A. K., Dalsgaard, S., Langley, K., Smith, G. D., . . . Thapar, A. (2016). Association of Genetic Risk Variants With Attention-Deficit/Hyperactivity Disorder Trajectories in the General Population. *JAMA Psychiatry, 73*(12), 1285-1292. doi:10.1001/jamapsychiatry.2016.2817

2575730 [pii]

Ritsner, M. S. (Ed.) (2009). *Neuropsychological Endophenotypes and Biomarkers* (Vol. 1): Springer Netherlands.

Scassellati, C., Bonvicini, C., Faraone, S. V., & Gennarelli, M. (2012). Biomarkers and attention-deficit/hyperactivity disorder: a systematic review and meta-analyses. *J Am Acad Child Adolesc Psychiatry, 51*(10), 1003-1019 e1020. doi:S0890-8567(12)00605-3 [pii]

10.1016/j.jaac.2012.08.015

Schwarz, G. (1978). Estimating the Dimension of a Model. *Ann. Statist., 6*(2), 461-464. doi:10.1214/aos/1176344136

Seidman, L. J., Valera, E. M., & Makris, N. (2005). Structural brain imaging of attention-deficit/hyperactivity disorder. *Biol Psychiatry, 57*(11), 1263-1272. doi:10.1016/j.biopsych.2004.11.019

Seidman, L. J., Valera, E. M., Makris, N., Monuteaux, M. C., Boriel, D. L., Kelkar, K., . . . Biederman, J. (2006). Dorsolateral prefrontal and anterior cingulate cortex volumetric abnormalities in adults with attention-deficit/hyperactivity disorder identified by magnetic resonance imaging. *Biol Psychiatry, 60*(10), 1071-1080. doi:10.1016/j.biopsych.2006.04.031

Semrud-Clikeman, M., Hooper, S. R., Hynd, G. W., Hern, K., Presley, R., & Watson, T. (1996). Prediction of group membership in developmental dyslexia, attention deficit hyperactivity disorder, and normal controls using brain morphometric analysis of magnetic resonance imaging. *Arch Clin Neuropsychol, 11*(6), 521-528. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/14588457

Shaw, P., De Rossi, P., Watson, B., Wharton, A., Greenstein, D., Raznahan, A., . . . Chakravarty, M. M. (2014). Mapping the development of the basal ganglia in children with attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry, 53*(7), 780-789 e711. doi:10.1016/j.jaac.2014.05.003

Shaw, P., Lerch, J., Greenstein, D., Sharp, W., Clasen, L., Evans, A., . . . Rapoport, J. (2006). Longitudinal mapping of cortical thickness and clinical outcome in children and adolescents with attention-deficit/hyperactivity disorder. *Arch Gen Psychiatry, 63*(5), 540-549. doi:10.1001/archpsyc.63.5.540

Sidhu, G. S., Asgarian, N., Greiner, R., & Brown, M. R. (2012). Kernel Principal Component Analysis for dimensionality reduction in fMRI-based diagnosis of ADHD. *Front Syst Neurosci, 6*, 74. doi:10.3389/fnsys.2012.00074

Smith, A. B., Taylor, E., Brammer, M., Toone, B., & Rubia, K. (2006). Task-specific hypoactivation in prefrontal and temporoparietal brain regions during motor inhibition and task switching in medication-naive children and adolescents with attention deficit hyperactivity disorder. *Am J Psychiatry, 163*(6), 1044-1051. doi:10.1176/ajp.2006.163.6.1044

Snyder, S. M., Rugino, T. A., Hornig, M., & Stein, M. A. (2015). Integration of an EEG biomarker with a clinician's ADHD evaluation. *Brain and Behavior*. doi:10.1002/brb3.330

Sole Puig, M., Perez Zapata, L., Puigcerver, L., Esperalba Iglesias, N., Sanchez Garcia, C., Romeo, A., . . . Super, H. (2015). Attention-Related Eye Vergence Measured in Children with Attention Deficit Hyperactivity Disorder. *PLoS One, 10*(12), e0145281. doi:10.1371/journal.pone.0145281

PONE-D-15-17294 [pii]

StataCorp. (2019). Stata Statistical Software: Release 16. College Station, TX: StataCorp LP.

The Express Scripts Lab. (2014). *Turning Attention to ADHD: U.S. Medication Trends for Attention Deficit Hyperactivity Disorder*. Retrieved from http://lab.express-scripts.com/insights/industry-updates/report-turning-attention-to-adhd

Thome, J., Ehlis, A. C., Fallgatter, A. J., Krauel, K., Lange, K. W., Riederer, P., . . . Gerlach, M. (2012). Biomarkers for attention-deficit/hyperactivity disorder (ADHD). A consensus report of the WFSBP task force on biological markers and the World Federation of ADHD. *World J Biol Psychiatry, 13*(5), 379-400. doi:10.3109/15622975.2012.690535

Thompson, P. M., Andreassen, O. A., Arias-Vasquez, A., Bearden, C. E., Boedhoe, P. S., Brouwer, R. M., . . . Ye, J. (2017). ENIGMA and the individual: Predicting factors that affect the brain in 35 countries worldwide. *NeuroImage, 145*, 389-408. doi:https://doi.org/10.1016/j.neuroimage.2015.11.057

Thompson, P. M., Jahanshad, N., Ching, C. R. K., Salminen, L. E., Thomopoulos, S. I., Bright, J., . . . Consortium, E. (2020). ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. *Transl Psychiatry, 10*(1), 100. doi:10.1038/s41398-020-0705-1

Thompson, P. M., Stein, J. L., Medland, S. E., Hibar, D. P., Vasquez, A. A., Renteria, M. E., . . . Alzheimer's Disease Neuroimaging Initiative, E. C. I. C. S. Y. S. G. (2014). The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav, 8*(2), 153-182. doi:10.1007/s11682-013-9269-5

Tian, L., Jiang, T., Wang, Y., Zang, Y., He, Y., Liang, M., . . . Zhuo, Y. (2006). Altered resting-state functional connectivity patterns of anterior cingulate cortex in adolescents with attention deficit hyperactivity disorder. *Neurosci Lett, 400*(1-2), 39-43. doi:10.1016/j.neulet.2006.02.022

Toyonaga, T., Shiga, T., Hirata, K., Yamaguchi, S., Takeuchi, W., Kudo, K., . . . Tamaki, N. (2017). Convolutional neural network (CNN) of MRI and FDG-PET images may predict hypoxia in glioblastoma. *Journal of Nuclear Medicine, 58*(supplement 1), 699. Retrieved from http://jnm.snmjournals.org/content/58/supplement_1/699.abstract

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS One, 14*(11), e0224365. doi:10.1371/journal.pone.0224365

Valera, E. M., Faraone, S. V., Murray, K. E., & Seidman, L. J. (2007). Meta-analysis of structural imaging findings in attention-deficit/hyperactivity disorder. *Biol Psychiatry, 61*(12), 1361-1369. doi:S0006-3223(06)00803-1 [pii]

10.1016/j.biopsych.2006.06.011

Visser, S. N., Danielson, M. L., Bitsko, R. H., Holbrook, J. R., Kogan, M. D., Ghandour, R. M., . . . Blumberg, S. J. (2014). Trends in the parent-report of health care provider-diagnosed and medicated attention-deficit/hyperactivity disorder: United States, 2003-2011. *J Am Acad Child Adolesc Psychiatry, 53*(1), 34-46 e32. doi:10.1016/j.jaac.2013.09.001

Wang, X., Jiao, Y., Tang, T., Wang, H., & Lu, Z. (2013). Altered regional homogeneity patterns in adults with attention-deficit hyperactivity disorder. *Eur J Radiol, 82*(9), 1552-1557. doi:10.1016/j.ejrad.2013.04.009

Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B., & Marquand, A. F. (2015). From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci Biobehav Rev, 57*, 328-349. doi:10.1016/j.neubiorev.2015.08.001

Yao, D., Guo, X., Zhao, Q., Liu, L., Cao, Q., Wang, Y., . . . Sui, J. (2018). Discriminating ADHD From Healthy Controls Using a Novel Feature Selection Method Based on Relative Importance and Ensemble Learning. *Conf Proc IEEE Eng Med Biol Soc, 2018*, 4632-4635. doi:10.1109/EMBC.2018.8513155

Zhang-James, Y., Chen, Q., Kuja-Halkola, R., Lichtenstein, P., Larsson, H., & Faraone, S. V. (2020). Machine-Learning prediction of comorbid substance use disorders in ADHD youth using Swedish registry data. *J Child Psychol Psychiatry*. doi:10.1111/jcpp.13226

Zhu, C. Z., Zang, Y. F., Cao, Q. J., Yan, C. G., He, Y., Jiang, T. Z., . . . Wang, Y. F. (2008). Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder. *Neuroimage, 40*(1), 110-120. doi:10.1016/j.neuroimage.2007.11.029

Zhu, C. Z., Zang, Y. F., Liang, M., Tian, L. X., He, Y., Li, X. B., . . . Jiang, T. Z. (2005). Discriminative analysis of brain function at resting-state for attention-deficit/hyperactivity disorder. *Med Image Comput Comput Assist Interv, 8*(Pt 2), 468-475. doi:10.1007/11566489_58

Zou, L., Zheng, J., Miao, C., Mckeown, M. J., & Wang, Z. J. (2017). 3D CNN Based Automatic Diagnosis of Attention Deficit Hyperactivity Disorder Using Functional and Structural MRI. *IEEE Access, 5*, 23626-23636. doi:10.1109/ACCESS.2017.2762703

Figure Captions

Figure 1. Best prediction accuracies reported in each study for each type of the available tests: K-Fold-CV, LOOCV or held-out tests.

Figure 2. Accuracy in studies published over the years.

Figure 3. Accuracy vs training sample size.

Sample size <300 were labeled as triangle and >300 are labeled as circle. The fitted line between accuracy and sample size were plotted for each test type.

Table 1

Table 1. Machine Learning Literature on ADHD Neuroimaging Data.

Study	Training Sample Size)	ADHD% (Training Set)	Test Sample Size	ADHD% (Test Set)	Data Source	Ages	Sex	Model	Features	Performance Metrics	Test Type	Accuracy	References	PMID/Conference
Aradhya, 2019	371	n.a	94	n.a	ADHD-200 subset (right-handed males)	Children and young adults (7-21)	M, F	CNN	rs-fMRI	Accuracy	K-Fold-CV(K=10)	70%	(Aradhya, Joglekar et al. 2019)	n.a
Ariyarathne, 2020	26	n.a	16	n.a	ADHD-200 subset	Children and young adults (8-21)	M, F	CNN	rs-fMRI	Accuracy, Sensitivity, Specificity	Held-out Test	85%	(Ariyarathne, Silva et al. 2020)	n.a
Bohland, 2012	776	37%	171	45%	ADHD-200	Children and young adults (7-21)	M, F	SVM	sMRI and rs-fMRI	Accuracy, AUC	K-Fold-CV(K=2) Held-out Test	74% 67%	(Bohland, Saperstein et al. 2012)	23267318
Brown MR 2012	668	36%	171	45%	ADHD-200	Children and young adults (7-21)	M, F	SVM	rs-fMRI	Accuracy	Held-out Test K-Fold-CV(K=10)	55% 71%	(Brown, Sidhu et al. 2012)	23060754
Chaim-Avancini, 2017	96	54%	n.a	n.a	Clinic and commmunity	Adults (18-50)	M, F	SVM	sMRI and DTI	Accuracy, ROC AUC, Sensitivity, Specificity, PPV, NPV	K-Fold-CV(K=10)	74%	(Chaim-Avancini, Doshi et al. 2017)	29080396
Chen, 2020	633	43%	n.a	n.a	ADHD-200	Children and young adults (7-21)	M, F	BHT	rs-fMRI	Accuracy, Sensitivity, Specificity	LOOCV	88%	(Chen, Tang et al. 2020)	32143793
Cheng W, 2012	239	41%	n.a	n.a	ADHD-200	Children and young adults (7-21)	M, F	SVM	rs-fMRI	Accuracy, Sensitivity, Specificity	LOOCV	76%	(Cheng, Ji et al. 2012)	22888314
Colby JB, 2012	776	37%	197	n.a	ADHD-200	Children and young adults (7-21)	M, F	SVM	sMRI and rs-fMRI	Accuracy, Sensitivity, Specificity	Held-out Test	59%	(Colby, Rudie et al. 2012)	22912605
Dai D, 2012	624	36%	n.a	n.a	ADHD-200	Children and young adults (7-21)	M, F	MKL	sMRI and rs-fMRI	Accuracy, Sensitivity, Specificity, J-statistic, F1-score, ROC AUC	K-Fold-CV(K=10) Held-out Test	68% 62%	(Dai, Wang et al. 2012)	22969710
Deshpande G, 2015	1177	37%	n.a	n.a	ADHD-200	Children and young adults (7-21)	M, F	FCCANN	rs-fMRI	Accuracy	LOOCV	90%	(Deshpande, Wang et al. 2015)	25576588
Dey, 2014	776	37%	n.a	n.a	ADHD-200	Children and young adults (7-21)	M, F	SVM	rs-fMRI	Accuracy, Sensitivity, Specificity	Training samples Held-out Test	71% 74%	(Dey, Rao et al. 2014)	24982615
Du J, 2016	216	55%	n.a	n.a	ADHD-200	Children and young adults (7-21)	M, F	SVM	rs-fMRI	Accuracy, Sensitivity, Specificity, ROC AUC	K-Fold-CV(K=10)	95%	(Du, Wang et al. 2016)	27166430
Eloyan A, 2012	776	37%	194	n.a	ADHD-200	Children and young adults (7-21)	M, F	Ensemble	sMRI and rs-fMRI and demographics	Accuracy, Sensitivity, Specificity	Held-out Test K-Fold-CV (n=184 randomly chosen internal test set)	61% 78%	(Eloyan, Muschelli et al. 2012)	22969709
Fair DA 2013	104	50%	n.a	n.a	ADHD-200	Children and young adults (7-21)	M, F	SVM	rs-fMRI	Accuracy, Sensitivity, Specificity	LOOCV	83%	(Fair, Nigg et al. 2012)	23382713
Ghiassian S, 2016	769	36%	171	45%	ADHD-200	Children and young adults (7-21)	M, F	MHPC	sMRI and rs-fMRI and demographics	Accuracy, Sensitivity, Specificity	Held-out Test	70%	(Ghiassian, Greiner et al. 2016)	28030565
Hao A, 2015	216	55%	41	71%	ADHD-200 (NYU subset)	Children and young adults (7-21)	M, F	DBN	fMRI	Accuracy, Sensitivity, Specificity	Held-out Test	49%	(Hao, He et al. 2015)	n.a
	85	28%	50	46%	ADHD-200 (Peking subset)	Children and young adults (7-21)	M, F	DBN	fMRI	Accuracy, Sensitivity, Specificity	Held-out Test	54%		
	83	27%	11	27%	ADHD-200 (KKI subset)	Children and young adults (7-21)	M, F	DBN	fMRI	Accuracy, Sensitivity, Specificity	Held-out Test	72%		
Hart H, 2014	60	50%	n.a	n.a	clinic and local community	Children and adolescents (10-17)	M, F	GPC	task-fMRI	Accuracy, ROC AUC, Sensitivity, Specificity, PPV, NPV	LOOCV	77%	(Hart, Chantiluke et al. 2014)	24123508
Iannaccone R, 2015	40	50%	n.a	n.a	Outpatient clinic and local schools	Adolescents (12-16)	M, F	SVM	task-fMRI	Accuracy, ROC AUC, sensitivity, Specificity	LOOCV	78%	(Iannaccone, Hauser et al. 2015)	25613588
Igual L, 2012	78	50%	n.a	n.a	URNC database	Children and adolescents (6-18)	M, F	SVM	sMRI of the caudate nucleus	Accuracy, Sensitivity, Specificity	K-Fold-CV(K=5)	73%	(Igual, Soliva et al. 2012)	22959658
Jie, 2016	216	55%	n.a	n.a	ADHD-200 (NYU subset)	Children and young adults (7-21)	M, F	SVM	rs-fMRI	Accuracy, ROC AUC, sensitivity, Specificity	LOOCV	83%	(Jie, Wee et al. 2016)	27060621
Johnston BA, 2014	68	50%	n.a	n.a	clinic and local schools	Children and adolescents (8-17)	M	SVM	sMRI	Accuracy, Sensitivity, Specificity	LOOCV	93%	(Johnston, Mwangi et al. 2014)	24819333
Kuang D, 2014	83	n.a	11	n.a	ADHD-200 (KKI subset)	Children and young adults (7-21)	M, F	DBM	rs-fMRI	Accuracy, Sensitivity, Specificity	Held-out Test	73%	(Kuang, Guo et al. 2014)	n.a
	85	28%	50	46%	ADHD-200 (Peking subset)	Children and young adults (7-21)	M, F				Held-out Test	54%		
	222	n.a	41	n.a	ADHD-200 (NYU subset)	Children and young adults (7-21)	M, F				Held-out Test	37%		
Lanka, 2019	759	37%	171	45%	ADHD-200	Children and young adults (7-21)	M, F	ensemble and ELM	rs-fMRI	Balanced Accuracy	Held-out Test	61%	(Lanka, Rangaprakash et al. 2019)	31691160
Lim L, 2013	48	60%	n.a	n.a	Clinic	Children and adolescents (10-18)	M	GPC	sMRI	Accuracy, AUC, Sensitivity, Specificity, PPV, NPV	LOOCV	79%	(Lim, Marquand et al. 2013)	23696841
Mao, 2019	626	46%	162	45%	ADHD-200	Children and young adults (7-21)	M, F	4D CNN	rs-fMRI	Accuracy, ROC AUC	Held-out Test	71%	(Mao, Su et al. 2019)	n.a
Olivetti E, 2012	923	38%	n.a	n.a	ADHD-200	Children and young adults (7-21)	M, F	ERT	rs-fMRI	Accuracy, FP, FN, TP, TN, Log(B10)	K-Fold-CV(K=10)	66%	(Olivetti, Greiner et al. 2012)	23060755
Olivetti E, 2015	923	38%	n.a	n.a	ADHD-200	Children and young adults (7-21)	M, F	ERT	rs-fMRI	Accuracy, MCC, J-statistic, F1-score, Log(B10)	K-Fold-CV(K=10)	62%	(Olivetti, Greiner et al. 2012)	27747500
Peng X, 2013	110	50%	n.a	n.a	ADHD-200 (Peking subset)	Children and young adults (7-21)	M, F	ELM	sMRI	Accuracy, ROC AUC	LOOCV	90%	(Peng, Lin et al. 2013)	24260229
Qureshi MN, 2016	106	50%	n.a	n.a	ADHD-200	Children and young adults (7-21)	M, F	H-ELM	sMRI	Accuracy	K-Fold-CV(K=10) K-Fold-CV(70/30 split)	80% 85%	(Qureshi, Min et al. 2016)	27500640

Qureshi MN, 2017	106	50%	28	50%	ADHD-200	Children and young adults (7-21)	M, F	ELM	sMRI and rs-fMRI	Accuracy, Sensitivity, Specificity, F1-score, Precision, Recall	Held-out Test	93%	(Qureshi, Oh et al. 2017)	28420972
Riaz, 2017	464	52%	65	44%	ADHD-200 (NeuroImaging, NYU and Peking subset)	Children and young adults (7-21)	M, F	CNN and SVM	rs-fMRI and demorgraphic	Accuracy	Held-out Test	69%	(Riaz, Asad et al. 2017)	n.a
Riaz, 2018a	442	43%	n.a	n.a	ADHD-200 (NeuroImaging, KKI, NYU and Peking subset)	Children and young adults (7-21)	M, F	SVM	rs-fMRI and demorgraphic	Accuracy, ROC AUC, sensitivity, Specificity	LOOCV	87%	(Riaz, Asad et al. 2018)	29137838
Riaz, 2018b	226	54%	n.a	n.a	ADHD-200 (NYU subset)	Children and young adults (7-21)	M, F	CNN	rs-fMRI	Accuracy, Sensitivity, Specificity	Held-out Test	73%	(Riaz, Asad et al. 2018)	n.a
Sato, 2012	759	36%	171	45%	ADHD-200	Children and young adults (7-21)	M, F	AdaBoost	rs-fMRI	Sensitivity, Specificity, Balanced Accuracy	Held-out Test	55%	(Sato, Hoexter et al. 2012)	23015782
Semrud-Clikeman, 1996	20	50%	n.a	n.a	Clinic and commmunity	Children and adolescents (6-16)	M, F	PDA	sMRI	Accuracy	training samples	87%	(Semrud-Clikeman, Hooper et al. 1996)	14588457
Sen, 2018	776	37%	171	45%	ADHD-200	Children and young adults (7-21)	M, F	SVM	sMRI and rs-fMRI	Accuracy, Sensitivity, Specificity, J-statistic	Held-out Test	67%	(Sen, Borle et al. 2018)	29664902
Shao, 2018	50	36%	16	36%	ADHD-200 (KKI subset)	Children and young adults (7-21)	M, F	SVM	rs-fMRI	Accuracy, Sensitivity, Specificity, MCC	Held-out Test		(Shao, Xu et al, 2018)	30009990
Sidhu, 2012	668	36%	171	45%	ADHD-200	Children and young adults (7-21)	M, F	SVM	rs-fMRI	Accuracy	training samples	76%	(Sidhu, Asgarian et al. 2012)	23162439
											Held-out Test	67%		
Tan, 2017	215	54%	n.a	n.a	ADHD-200 (NYU subset)	Children and young adults (7-21)	M, F	SVM	sMRI and rs-fMRI	Accuracy, AUC, sensitivity, Specificity, Balanced Accuracy	K-Fold-CV(K=10)	68%	(Tan, Guo et al. 2017)	28943846
Tang, 2019	633	43%	n.a	n.a	ADHD-200	Children and young adults (7-21)	M, F	BHT	rs-fMRI	Accuracy, Sensitivity, Specificity	LOOCV	92%	(Tang, Wang et al. 2019)	30938224
Tang, 2020	633	43%	n.a	n.a	ADHD-200	Children and young adults (7-21)	M, F	BHT	rs-fMRI	Accuracy, Sensitivity, Specificity	LOOCV	98%	(Tang, Li et al. 2020)	n.a
Wang, 2013	46	50%	n.a	n.a	FCON_1000	Adults (18-50)	M, F	SVM	rs-fMRI	Accuracy, Sensitivity, Specificity	LOOCV	80%	(Wang, Jiao et al. 2013)	23684384
Wang, 2018	71	51%	n.a	n.a	ADHD-200 subset	Children and adolescents (6-18)	M, F	SVM	sMRI	Accuracy, Sensitivity, Specificity	LOOCV	75%	(Wang, Jiao et al. 2018)	30031733
Xiao, 2016	47	68%	n.a	n.a	clinic	n.a	n.a	Lasso	sMRI	Accuracy, Sensitivity, Specificity	LOOCV	81%	(Xiao, Bledsoe et al. 2016)	27747592
Yao, 2018	189	59%	n.a	n.a	clinic	Adults (18-34)	M, F	Ensemble	rs-fMRI	Accuracy, Sensitivity, Specificity	K-Fold-CV(K=10)	80%	(Yao, Guo et al. 2018)	30441383
						Children and adolescents (6-14)	M					86%		
Yoo, 2019	94	50%			Clinic	Children and adolescents (6-17)	M, F	RF	sMRI and rs-fMRI and DTI	Accuracy, AUC, Sensitivity, Specificity, PPV, NPV	LOOCV	85%	(Yoo, Kim et al. 2019)	31321662
											Held-out Test	69%		
Zhu CZ, 2008	20	45%	n.a	n.a	community	Children and adolescents (11-17)	M	FDA	rs-fMRI	Accuracy, Sensitivity, Specificity	LOOCV	85%	(Zhu, Zang et al. 2008)	18191584
Zou, 2017	559	35%	171	45%	ADHD-200	Children and young adults (7-21)	M, F	3D CNN	rs-fMRI and sMRI	Accuracy	Held-out Test	69%	(Zou, Zheng et al. 2017)	n.a
Zu, 2019	216	55%	n.a	n.a	ADHD-200 (NYU subset)	Children and young adults (7-21)	M, F	STM	rs-fMRI	Accuracy	K-Fold-CV(K=10)	65%	(Zu, Gao et al. 2019)	29948906

Note:
AUC, the area under the ROC curve (AUC) BHT,
Binary Hypothesis Testing
CNN, Convolutionary Neural Net;
CV, cross-validation; LOOCV, leave-one-out cross validation;
DBN, Deep Bayesian Network; DBM, Deep Belief Network;
ELM, extreme learning machine; H-ELM, hierarchical extreme learning machine.
ERT, extremely randomized tree
FCON_1000, 1000 Functional Connectomes Project database (http://www.nitrc.org/projects/fcon_1000)
FDA, Fisher discriminative analysis
fMRI, funcitonal MRI; rs-fMRI, resting state- functional MRI; sMRI, structure MRI; DTI, diffusion tensor imaging GBM,
a gradient boosting method;
GPC, Gaussian process classifiers;
Log(B_{10}), the log of the Bayes factor for the hypothesis of dependence vs. independence;
MCC, Matthew's correlation coefficient
MHPC, the histogram of oriented gradients (HOG)-feature-based patient classification;
MKL, multi-kernellearning; FCCANN, fully connected cascade artificial neural network;
PDA, predictive discriminant analysis
PPV, Positive predictive value; NPV, Negative predictive value
RF, Random Forest
SVM, support vector machine; STM, Support tensor machine.
TP, the number of true positive diagnosis; TN, the number of true negative diagnosis; FP, the number of false positive diagnosis; FN, and the number of false negative diagnosis.

*Balanced Accuracy = (sensitivity + specificity)/2

Figure 1

Figure 1.

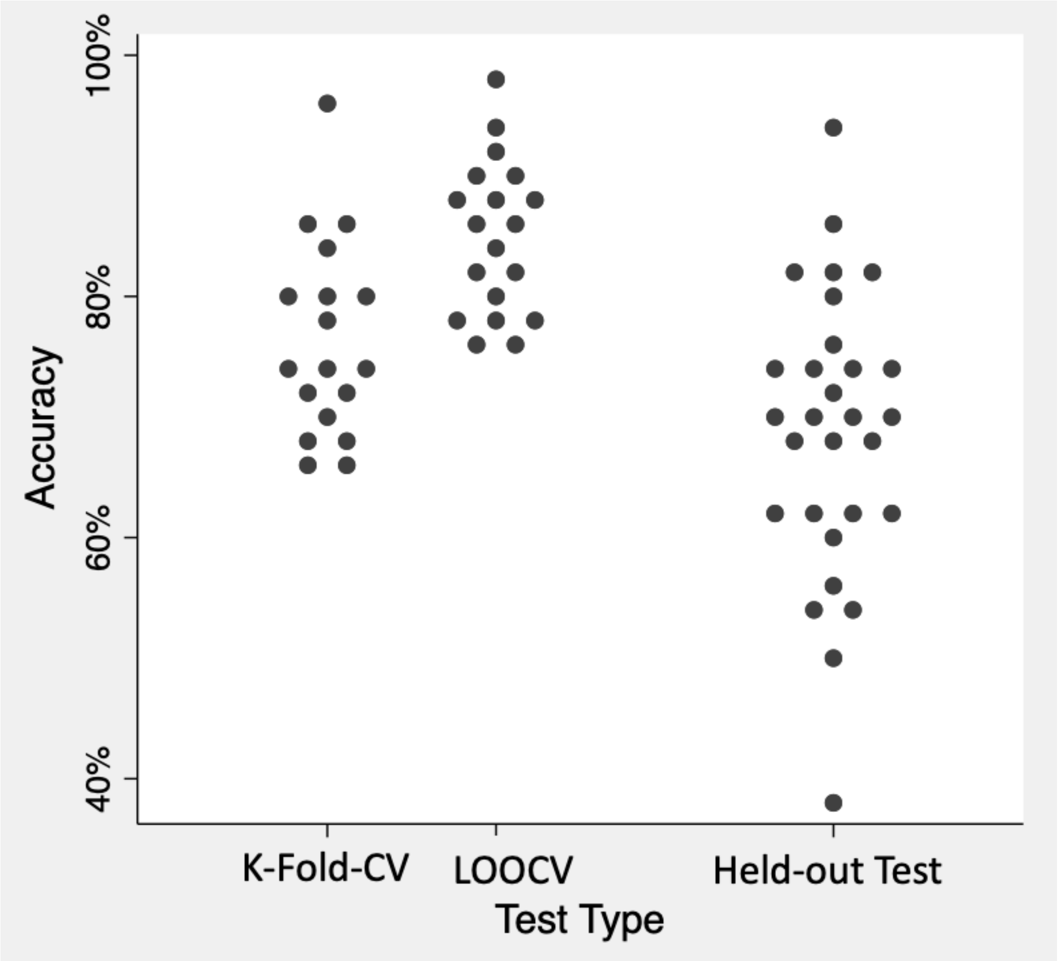


Figure 2

Figure 2.

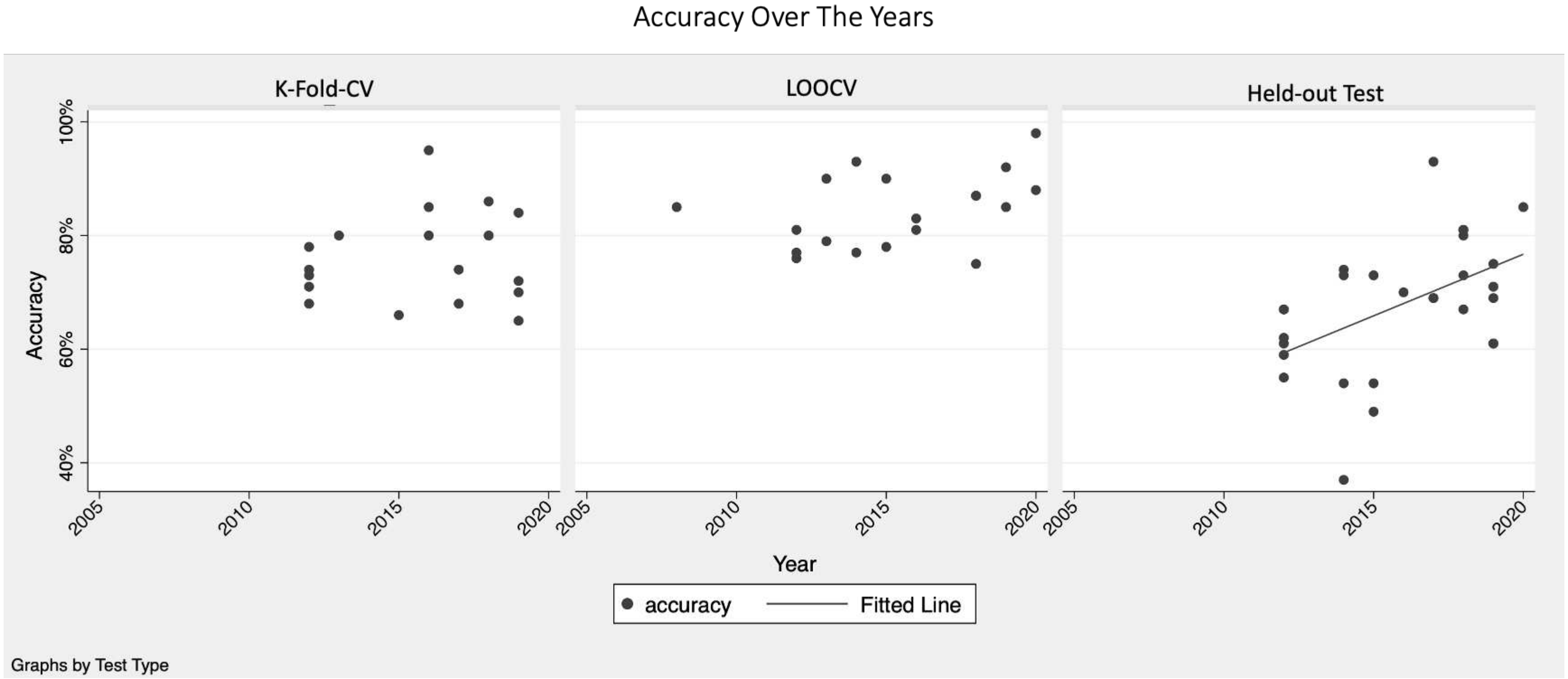
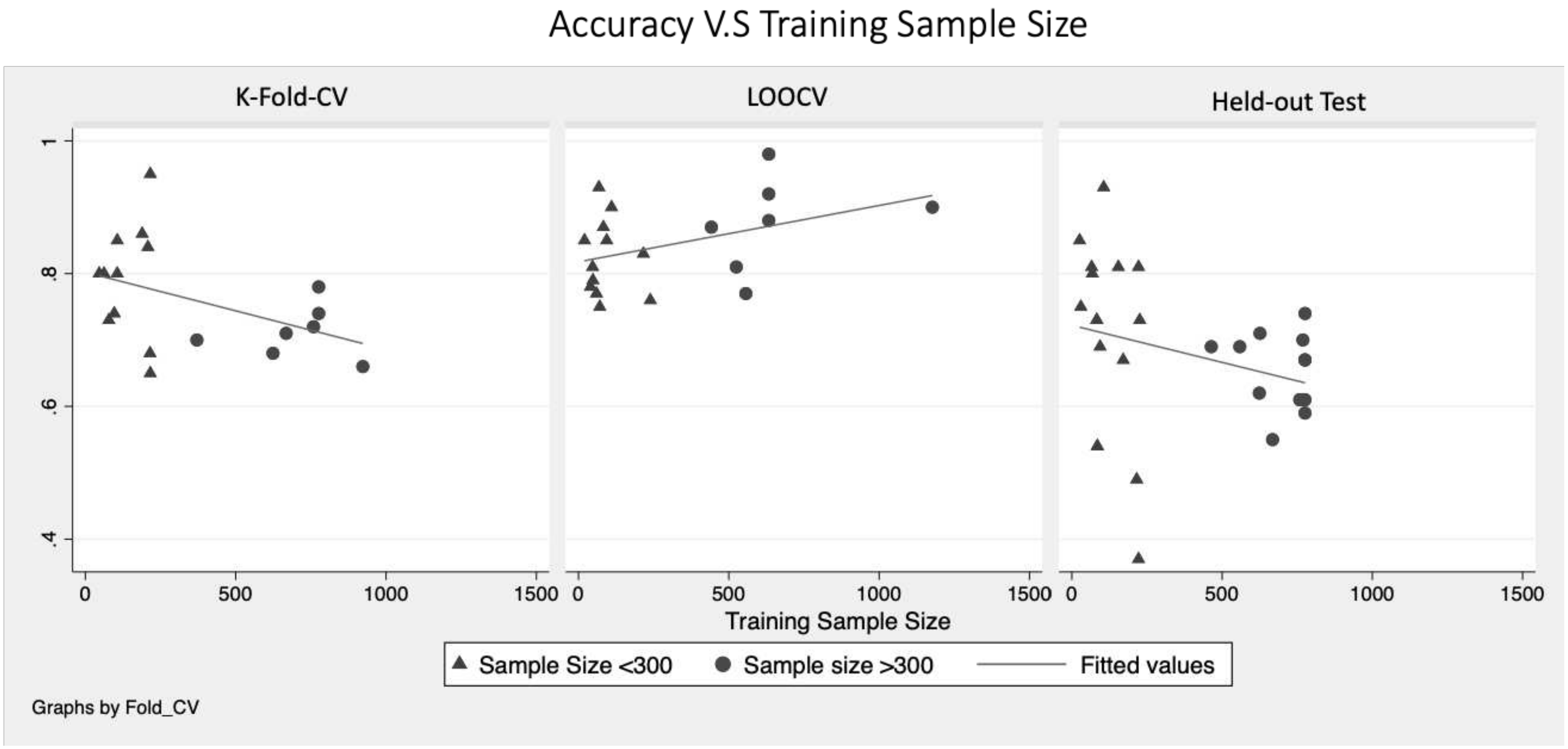
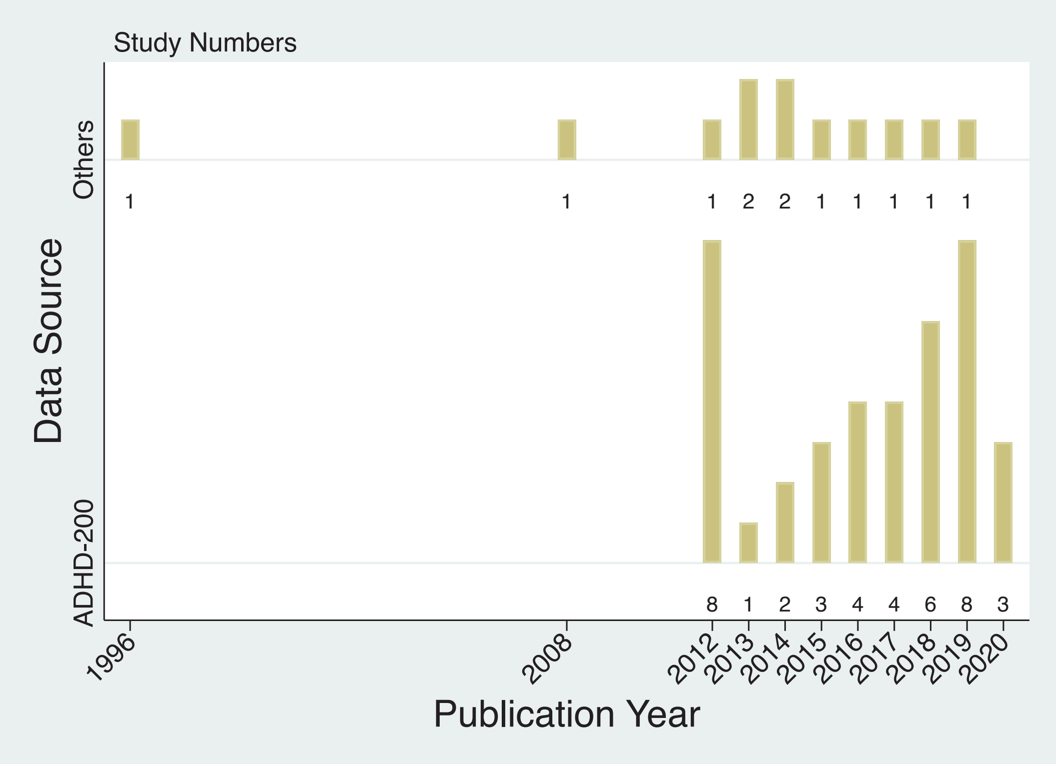


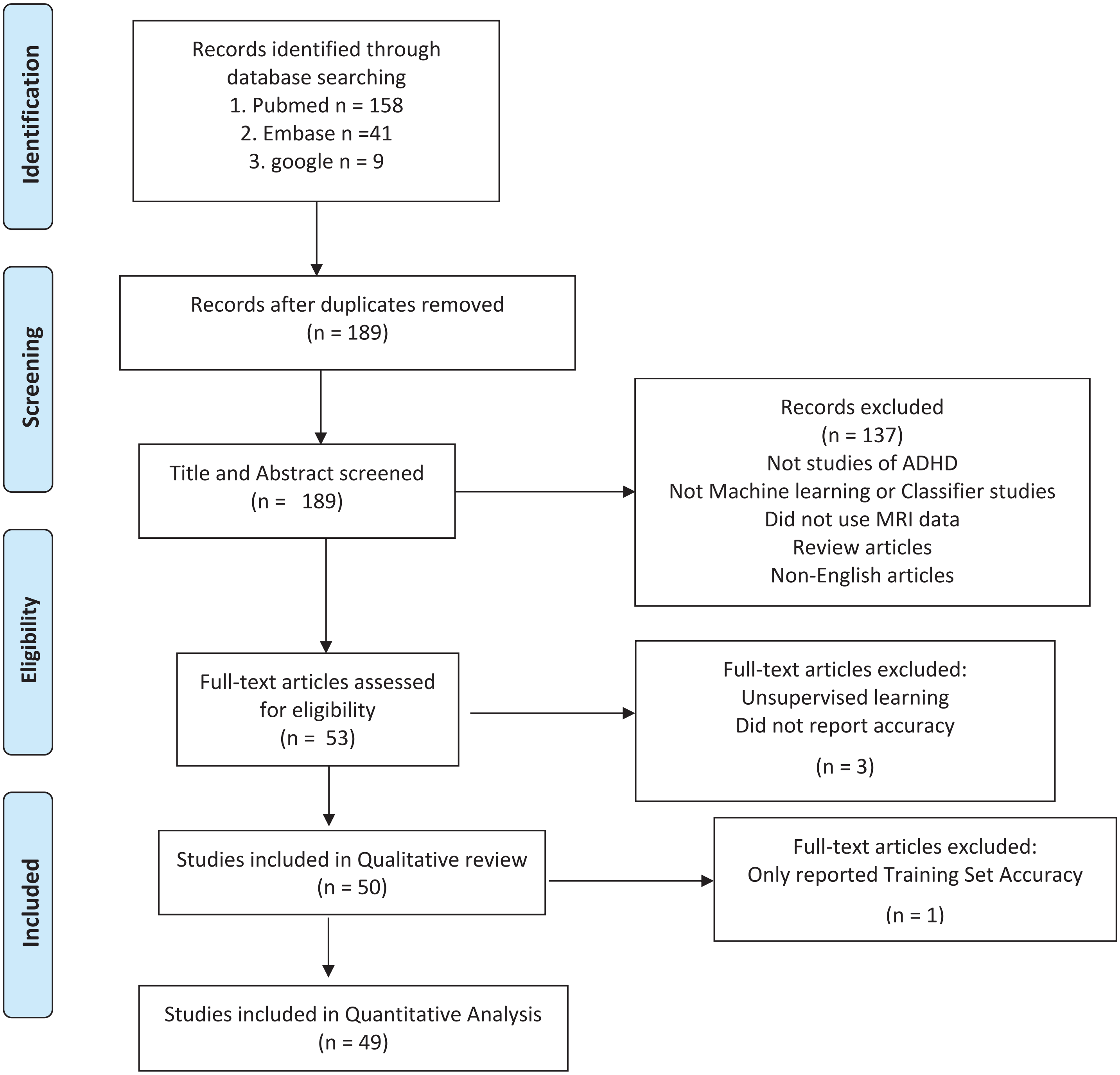
Figure 3.


Accuracy V.S Training Sample Size
K-Fold-CV
LOOCV
Held-out Test
.4
.6
.8
1
0
500
1000
1500
Training Sample Size
▲ Sample Size <300
● Sample size >300
Fitted values
Graphs by Fold_CV

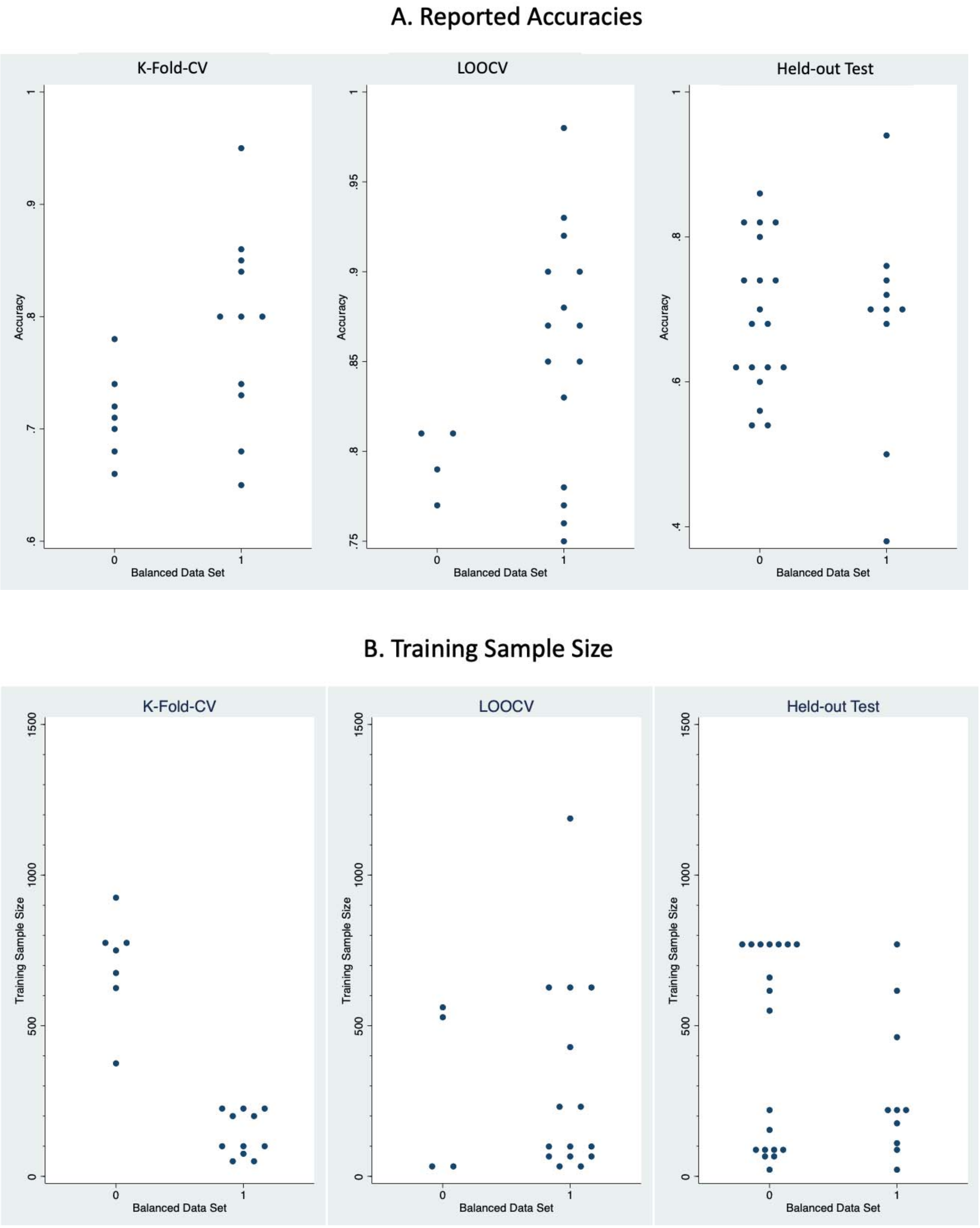
Supplementary Figure 1. Numbers of publication in each year. The top row includes studies that used non-ADHD-200 data; the bottom row includes studies that used ADHD-200 data. The numbers for each year are labeled beneath the bar.



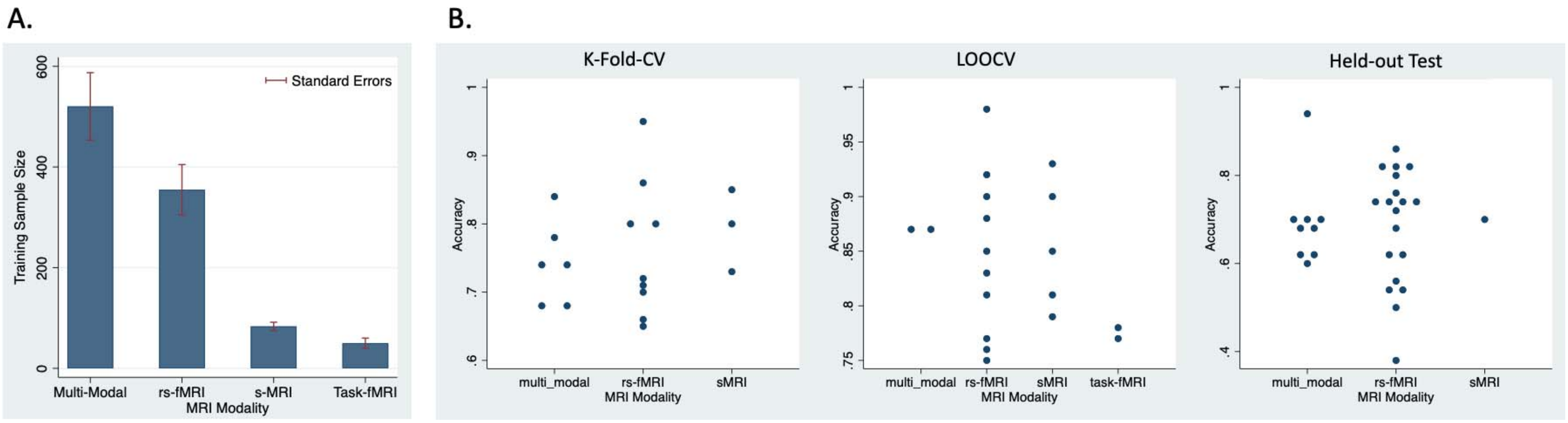
Supplementary Figure 2. PRISMA flow diagram for review and meta-analysis


Identification
Records identified through database searching
1. Pubmed n = 158
2. Embase n =41
3. google n = 9
Screening
Records after duplicates removed
(n = 189)
Title and Abstract screened
(n = 189)
Records excluded
(n = 137)
Not studies of ADHD
Not Machine learning or Classifier studies
Did not use MRI data
Review articles
Non-English articles
Eligibility
Full-text articles assessed for eligibility
(n = 53)
Full-text articles excluded:
Unsupervised learning
Did not report accuracy
(n = 3)
Included
Studies included in Qualitative review
(n = 50)
Full-text articles excluded:
Only reported Training Set Accuracy
(n = 1)
Studies included in Quantitative Analysis
(n = 49)

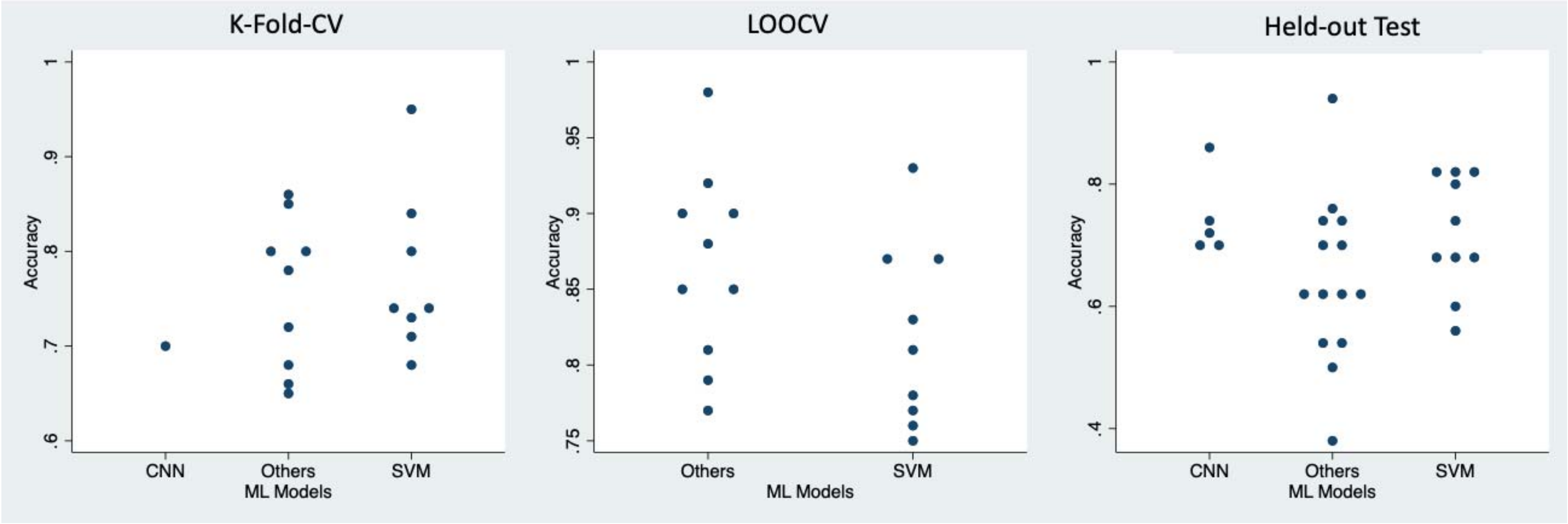
Supplementary Figure 3. A. Reported accuracies and training data balancing. B. Training sample sizes in balanced vs unbalanced studies.



Supplementary Figure 4. A. Mean and standard errors of the sample size for multi-modality, rs-fMRI, sMRI and task-based fMRI studies. B. Accuracies in studies using different MRI modalities.



Supplementary Figure 5. Accuracies in studies using different ML models.


K-Fold-CV
Accuracy
CNN
Others
SVM
ML Models
LOOCV
Accuracy
Others
SVM
ML Models
Held-out Test
Accuracy
CNN
Others
SVM
ML Models